Smart Innovation, Systems and Technologies 105

Suresh Chandra Satapathy
Vikrant Bhateja · Swagatam Das *Editors*

Smart Intelligent Computing and Applications

Proceedings of the Second International Conference on SCI 2018, Volume 2

KES International

Springer

Smart Innovation, Systems and Technologies

Volume 105

The Smart Innovation, Systems and Technologies book series encompasses the topics of knowledge, intelligence, innovation and sustainability. The aim of the series is to make available a platform for the publication of books on all aspects of single and multi-disciplinary research on these themes in order to make the latest results available in a readily-accessible form. Volumes on interdisciplinary research combining two or more of these areas is particularly sought.

The series covers systems and paradigms that employ knowledge and intelligence in a broad sense. Its scope is systems having embedded knowledge and intelligence, which may be applied to the solution of world problems in industry, the environment and the community. It also focusses on the knowledge-transfer methodologies and innovation strategies employed to make this happen effectively. The combination of intelligent systems tools and a broad range of applications introduces a need for a synergy of disciplines from science, technology, business and the humanities. The series will include conference proceedings, edited collections, monographs, handbooks, reference books, and other relevant types of book in areas of science and technology where smart systems and technologies can offer innovative solutions.

High quality content is an essential feature for all book proposals accepted for the series. It is expected that editors of all accepted volumes will ensure that contributions are subjected to an appropriate level of reviewing process and adhere to KES quality principles.

More information about this series at http://www.springer.com/series/8767

Suresh Chandra Satapathy · Vikrant Bhateja
Swagatam Das
Editors

Smart Intelligent Computing and Applications

Proceedings of the Second International Conference on SCI 2018, Volume 2

Springer

Editors
Suresh Chandra Satapathy
School of Computer Engineering
KIIT Deemed to be University
Bhubaneswar, Odisha, India

Vikrant Bhateja
Department of Electronics and Communication Engineering
Shri Ramswaroop Memorial Group of Professional Colleges
Lucknow, Uttar Pradesh, India

Swagatam Das
Electronics and Communication Sciences Unit
Indian Statistical Institute
Kolkata, West Bengal, India

ISSN 2190-3018 ISSN 2190-3026 (electronic)
Smart Innovation, Systems and Technologies
ISBN 978-981-13-1926-6 ISBN 978-981-13-1927-3 (eBook)
https://doi.org/10.1007/978-981-13-1927-3

Library of Congress Control Number: 2018950782

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Organizing Committee

Organizing Chairs

Dr. Suresh Chandra Satapathy, PVPSIT, India
Dr. B. V. SubbaRao, PVPSIT, India

Program Chairs

Dr. M. V. Rama Krishna, PVPSIT, India
Dr. J. Rajendra Prasad, PVPSIT, India

Program Co-chairs

Dr. S. Madhavi, PVPSIT, India
Dr. P. V. S. Lakshmi, PVPSIT, India
Dr. A. Sudhir Babu, PVPSIT, India

Conference Convenor

Dr. B. Janakiramaiah, PVPSIT, India

Conference Co-convenor

Dr. P. E. S. N. Krishna Prasad, PVPSIT, India
Mr. Y. Suresh, PVPSIT, India

Publicity Chairs

Ms. A. Ramana Lakshmi, PVPSIT, India
Mr. A. Vanamala Kumar, PVPSIT, India
Ms. D. Kavitha, PVPSIT, India

Special Session Chairs

Dr. B. Srinivasa Rao, PVPSIT, India
Ms. G. Reshma, PVPSIT, India

Workshop/Tutorial Chairs

Mr. B. N. Swamy, PVPSIT, India
Ms. A. Haritha, PVPSIT, India

Web Master

Mr. K. Syama Sundara Rao, PVPSIT, India
Mr. L. Ravi Kumar, PVPSIT, India

Organizing Committee Members

Computer Science and Engineering

Ms. J. Rama Devi
Ms. G. Lalitha Kumari
Ms. V. Siva Parvathi
Ms. B. Lakshmi Ramani
Mr. I. M. V. Krishna
Ms. D. Swapna
Mr. K. Vijay Kumar
Ms. M. Sailaja
Ms. Y. Surekha
Ms. D. Sree Lakshmi
Ms. T. Sri Lakshmi
Mr. N. Venkata Ramana Gupta
Ms. A. Madhuri
Ms. Ch. Ratna Jyothi
Mr. P. Anil Kumar
Mr. S. Phani Praveen
Mr. B. Vishnu Vardhan
Mr. D. Lokesh Sai Kumar
Ms. A. Divya
Mr. A. Yuva Krishna
Mr. Ch. Chandra Mohan
Mr. K. Venkatesh
Mr. V. Rajesh
Mr. M. Ramgopal

Information Technology

Ms. J. Sirisha
Mr. K. Pavan Kumar
Ms. G. Lakshmi
Mr. D. Ratnam
Mr. M. Sundara Babu

Mr. S. Sai Kumar
Ms. K. Swarupa Rani
Mr. P. Ravi Prakash
Mr. T. D. Ravi Kiran
Ms. Y. Padma
Ms. K. Sri Vijaya
Ms. D. Leela Dharani
Mr. G. Venugopal
Ms. M. Sowjanya
Mr. K. Prudviraju
Mr. R. Vijay Kumar Reddy

Preface

The Second International Conference on Smart Computing and Informatics (SCI 2018) was successfully organized by Prasad V. Potluri Siddhartha Institute of Technology, Vijayawada, Andhra Pradesh. The objective of this international conference was to provide a platform for academicians, researchers, scientists, professionals, and students to share their knowledge and expertise in the field of smart computing comprising of soft computing, evolutionary algorithms, swarm intelligence, Internet of things, machine learning, etc. and address various issues to increase an awareness of technological innovations and to identify challenges and opportunities to promote the development of multidisciplinary problem-solving techniques and applications. Research submissions in various advanced technology areas were received, and after a rigorous peer review process with the help of technical program committee members, elite quality papers were accepted. The conference featured special sessions on various cutting-edge technologies which were chaired by eminent professors. Many distinguished researchers like Dr. Lakhmi C. Jain, Australia; Dr. Nabil Khelifi, Springer, Germany; Dr. Roman Senkerik, Tomas Bata University in Zlin, Czech Republic; Dr. B. K. Panigrahi, IIT Delhi; and Dr. S. Das, Indian Statistical Institute, Kolkata, attended the conference and delivered the talks.

Our sincere thanks to all special session chairs, reviewers, authors for their excellent support. We would like to thank Siddhartha Academy for all the support to make this event possible. Special thanks to Sri B. Sreeramulu, Convenor, PVPSIT, and Dr. K. Sivaji Babu, Principal, PVPSIT, for their continuous support. We would like to extend our special thanks to very competitive team members from the Department of CSE and IT, PVPSIT, for successfully organizing the event.

	Editorial Boards
Bhubaneswar, India	Dr. Suresh Chandra Satapathy
Lucknow, India	Dr. Vikrant Bhateja
Kolkata, India	Dr. Swagatam Das

Details of Special Sessions

1. Special Session on "Emerging Trends in BigData, IoT and Social Networks (ETBIS)"

Session Chair

Dr. L. D. Dhinesh Babu, VIT University, Vellore, TN, India

Co-session Chair(s)

Dr. S. Sumathy, VIT University, Vellore, TN, India
Dr. J. Kamalakannan, VIT University, Vellore, TN, India

2. Special Session on "Applications of Computing in Interdisciplinary Domains"

Session Chair

Prof. Dr. V. Suma
Dean, Research and Industry Incubation Centre
Professor, Department of Information Science and Engineering,
Dayananda Sagar College of Engineering
Bengaluru, 560078, India

3. "Recent Trends in Data Science and Security Analytics"

Session Chair(s)

Dr. Sireesha Rodda, GITAM University, Visakhapatnam, India
Dr. Hyma Janapana, GITAM University, Visakhapatnam, India

4. Special Session on "Trends in Smart Healthcare Technologies (TSHT)"

Session Chair

Dr. K. Govinda, VIT University, Vellore, TN, India

Co-session Chair(s)

Dr. R. Rajkumar, VIT University, Vellore, TN, India
Dr. N. Sureshkumar, VIT University, Vellore, TN, India

Contents

About the Editors

Dr. Suresh Chandra Satapathy is currently working as Professor, School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar, India. He obtained his Ph.D. in computer science and engineering from JNTU, Hyderabad, and M.Tech. in CSE from NIT Rourkela, Odisha, India. He has 27 years of teaching experience. His research interests are data mining, machine intelligence, and swarm intelligence. He has acted as program chair of many international conferences and edited six volumes of proceedings from Springer LNCS and AISC series. He is currently guiding eight scholars for Ph.D. He is also Senior Member of IEEE.

Dr. Vikrant Bhateja is Professor, Department of Electronics and Communication Engineering, Shri Ramswaroop Memorial Group of Professional Colleges (SRMGPC), Lucknow, and also Head (Academics and Quality Control) in the same college. His areas of research include digital image and video processing, computer vision, medical imaging, machine learning, pattern analysis and recognition, neural networks, soft computing, and bio-inspired computing techniques. He has more than 90 quality publications in various international journals and conference proceedings. He has been on TPC and chaired various sessions from the above domain in international conferences of IEEE and Springer. He has been Track Chair and served in the core-technical/editorial teams for international conferences: FICTA 2014, CSI 2014, and INDIA 2015 under Springer ASIC series; INDIACom-2015 and ICACCI 2015 under IEEE. He is Associate Editor in *International Journal of Convergence Computing* (IJConvC) and also serving in the editorial board of *International Journal of Image Mining* (IJIM) under Inderscience Publishers. At present, he is Guest Editor for two special issues floated in *International Journal of Rough Sets and Data Analysis* (IJRSDA) and *International Journal of System Dynamics Applications* (IJSDA) under IGI Global publications.

Dr. Swagatam Das received B.E. Tel.E., M.E. Tel.E. (Control Engineering), and Ph.D., all from Jadavpur University, India, in 2003, 2005, and 2009, respectively. He is currently serving as Assistant Professor in the Department of Electronics and Communication Sciences at the Indian Statistical Institute, Kolkata, India. His research interests include evolutionary computing, pattern recognition, multiagent systems, and wireless communication. He has published one research monograph, one edited volume, and more than 200 research articles in peer-reviewed journals and international conferences.

An Adaptive ARP Approach for Fog-Based RSU Utilization

G. Jeya Shree and S. Padmavathi

Abstract The rapid growth in the development of IoT leads to the increased demand for real-time and low-latency services, which is challenging for the traditional cloud computing model. One best solution emerged, is the Fog Computing, which provides services and resources at the network's edge. In this paper, we consider one of the IoT applications in smart cities, Vehicular Monitorization. Every vehicle in the smart cities communicates their information through the Road Side Units (RSU). The RSU must be utilized optimally and minimally while covering up a maximum range of vehicles. This paper proposes an infrastructure for vehicles and RSUs to use the Fog computing layer for its enumeration process and proposes an efficient Advanced Resource Provisioning (ARP) algorithm to minimize the number of RSUs utilized in the vehicular monitorization. Results show that the cost and power consumption of RSUs has also been reduced considerably with the minimization of RSU's.

1 Introduction

The emerging trends in the Smart Vehicular monitorization is to use the Fog computing over cloud for its processing of IoT data such as traffic data, vehicle data, etc., Already Cloud computing has been deployed to serve as a solution to vehicles which require computational, storage and network resources. But the requirement such as mobility, location awareness, low latency, etc., demand fog computing which brings computations to the edge. The Road Side Unit (RSU) has an important role to play in the communication and transformation of messages in vehicular monitorization. Therefore, it is necessary to find an optimal location for placing the RSU [1].

G. Jeya Shree (✉) · S. Padmavathi
Department of Computer Science and Engineering, Thiagarajar College
of Engineering, Madurai 625015, India
e-mail: jaishreesha.28@gmail.com

S. Padmavathi
e-mail: spmcse@tce.edu

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_1

Fog computing is an emerging paradigm that extends cloud computing and services at the network edge. In contrast to cloud, fog computing is aimed at deploying services in a widely distributed manner whereas it is centralized in cloud [2]. Fog provides storage and computation resources as well as application services to the users, like cloud [3]. Fog computing is closely related to IoT. IoT is generating an enormous amount of data day by day. By the time the data reaches the cloud for its analysis, the actual need for that data might be lost [2]. Arkian et al. [4] proposes a MIST-Fog based analytics scheme for cost-efficient resource provisioning for IoT application. The data consumer association, task distribution and virtual machine placement problem is also considered in minimizing the overall cost while the QoS requirement is still satisfied [4]. A fog node can be any device with storage, computing and network connectivity. The mobile vehicles and infrastructures in roads are also fog nodes [5]. Fog layer acts on IoT data in a fraction of seconds and analyses the most time sensitive data at the edge close to the end-users where it is generated. This reduces the cost, time and effort in outsourcing them to the cloud. Fog has been emerged as a powerful platform to deploy various applications like energy, health care, traffic and so on [6]. Thus, fog computing outperforms cloud by creating a distributed framework for IoT to cope up with needs of sensors and embedded systems such as data processing and storage.

Considering the ITS applications in smart cities, there are various solutions available which aims at improving the overall infrastructure by improving RSU functionalities. One such solution is proposed by Sankaranarayanan and Mala [7] which uses genetic algorithm to model an Optimized RSU based Travel Time system to estimate the travel time for vehicular users. However, it is challenging to identify the optimal location for placing the RSUs as well as to minimize its cost of installation on roads. Nawaz et al. [8] proposed a modified CLB approach which estimates the number of RSUs a vehicle cross while heading its destination, thus the connectivity problem in vehicles had overcome. By this the overload of a RSU can be redistributed among other RSUs and the performance of the system can be maximized. By placing the RSUs using scalable Task Duplication Based (TDB) technique, the coverage can be maximized as well as the efficiency can be increased [9]. Simple geometric rules that depend on sending node's position is applied to extend the coverage area of RSU. The sending nodes are responsible for the propagation of data in different directions [10]. The Cooperative Load Balancing transfer (CLB) mechanism is available among RSUs, that distributes requests of a heavily loaded RSU to a lesser loaded neighbour RSU [11].

First, we consider Efficient ARP algorithm for minimizing the number of RSU's utilized in the smart vehicular monitorization while minimizing the cost and power consumption of RSU. Then, we propose fog computing layer for computation and enumeration of traffic data coming from the RSU's and connected vehicles. Finally, we consider the simulation of essential nodes (fog nodes, sensors, actuators) in the fog environment. Then we perform a comparative performance analysis before and after ARP.

The paper is organized as follows. In Sect. 2, we present the preliminaries and background of fog computing. The proposed scheme and its methodology are discussed in Sect. 3 and in its subsections. In Sect. 4, we discuss the results and performance analysis. Finally, Sect. 5 concludes the paper.

2 Preliminaries

2.1 Need for Fog Computing

Nowadays, the Internet of Things (IoT) has been adopted by a large number of organizations and enterprises. So, the demand for quick access of such enormous amounts of data is keep on increasing. The characterization of Big Data is along three dimensions such as volume, velocity and variety. But, IoT use cases like smart transportation, smart cities and smart grid are generally distributed in nature. Hence this needs a fourth dimension to the characterization of Big Data, Geo-distribution. Outsourcing all the operations to cloud, causes delay and network congestion that hinder the performance of cloud. Hence, a distributed intelligent platform for managing the distributed computing, networking and storage resources is needed at the edge. This is where the concept of 'Fog Computing' comes to play, which extends from the edge to the cloud, in a geographically distributed manner. So, instead of an increasingly backed up centralized data model (cloud), we will start to use a decentralization of data (fog). Fog computing is generally associated with cloud computing thus forming a three-layer model 'Things-Fog-Cloud'. The ideal use cases of fog computing are Agriculture, Wind Energy, Surveillances and Smart cities. In this paper, we consider the use case of smart city application only. Figure 1 shows the three-layer architectural model.

2.2 Intelligent Transportation System (ITS)

As per the definition given by Intelligent Transportation Society of America in the year 1998, the aim of Intelligent Transportation System (ITS) is to use technology in transportation to save lives, money and to improve safety as well as travel times on the transportation system. The major goals of an ITS system is to provide driving comfort and to reduce any kind of fatal and financial losses due to the road accidents. Some examples of systems concerned by ITS such as traffic management, Advance vehicle control and safety, Emergency management, Rail Road crossing safety and Electronic payment.

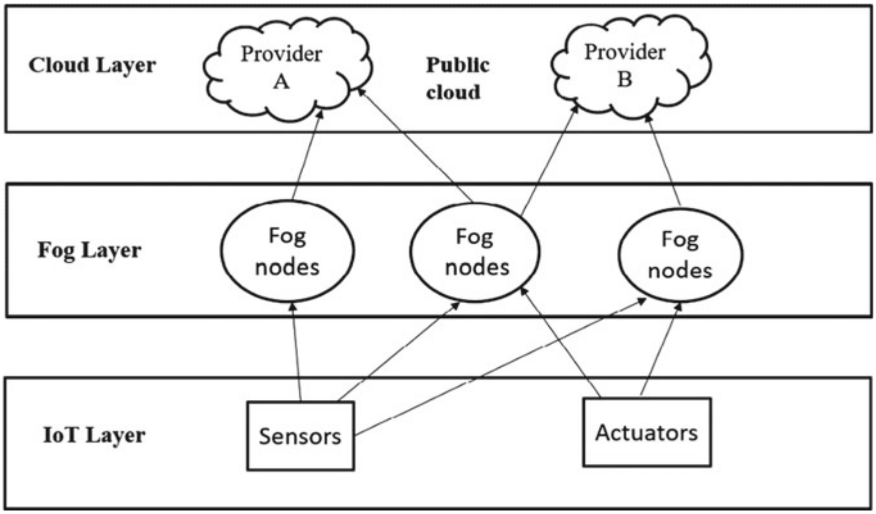


Fig. 1 'Things-Fog-Cloud' architecture

2.3 RSU—Road Side Units

In vehicular network, Vehicle-to-Vehicle communication (V2V) and Vehicle-to-Infrastructure communication (V2I) are the two main communication architectures. Here the vehicles are On-Board Units (OBU) and the infrastructures on the roads are Road Side Units (RSU). Road Side Unit is the Computing device located on the roadside that provides connectivity support to passing vehicles. The routing in vehicular networks can be improved by the capabilities of RSU, the range and the reliability of the V2I communications is increased due to the higher antenna heights. Considering the case of emergency messages, these characteristics are important in avoiding network congestion by load balancing the traffic.

This paper presents an algorithm for minimizing the number of RSU's that are provisioned in Intelligent Transportation Systems (ITS).

3 Proposed Methodology

The problem is formulated as follows. Given a set of 'm' lanes, $L = \{l_1, l_2, l_3\ldots, l_\mathrm{m}\}$ with 'μ' RSU's placed sporadically in some l_i's, where $0 \leq i \leq 1$ and $\mu \leq 1$ with each μ's storing 'λ' units of power and costing an amount 'k'. The task is to minimize μ, λ, and k such that,

$$\bigcup_{i=0}^{l} S_{\mu_i} \leq L \tag{1}$$

Equation (1) shows the objective of ARP algorithm, that is minimizing the number of RSUs utilized. The equation ensures that every lane is mapped with some RSUs.

3.1 ARP Algorithm for Minimizing RSU Utilization

For the optimal utilization of RSU over a given set of lanes several parameters are considered such as lane number, length of the lane, cost of installation of RSU and power consumed by the RSU. Several other parameters such as driving patterns, frequency of accidents, vehicle distribution over time are also considered to decide on where to place the RSU. i.e., On lanes where traffic and vehicle movement are less, the RSU in that lane can be set off and the nearby lane's RSU can take care of this lane, thus saving one RSU. Similarly, in this way we consider all the parameters and optimally utilize the minimum number of RSU's from the complete set of available RSU.

ARP Algorithm:

Input: 'm' number of lanes and 'm' binary codes representing the presence of RSU in the respective lanes.

Output: Minimization in the number of RSUs utilized.

Steps:

(1) Input the number of lanes 'm'
(2) Input 'm' binary codes for m lanes with '0' representing no RSUs in that lane and '1' representing the presence of RSU in that lane
(3) Considering the parameters, the less and heavy traffic lane is identified. Thus, the right locations for placing the RSU will be known
(4) If the less traffic lanes contain RSU in it, it can be either be set OFF to save its utilization or can be used to cover some additional adjacent neighbour lanes
(5) In heavy traffic lanes, RSU is set ON to its maximum coverage
(6) Similarly, every other parameter is considered for the optimal number of RSUs utilization

3.2 Proposed Infrastructure for Vehicle Monitorization

The proposed infrastructure for RSU and vehicles uses fog computing for its transformation of data where the enumeration and processing of data can be carried out easily and more effectively. With more amount of data transferring to and from the cloud causes network traffic and the network traffic will lead to congestion and increase in latency. The benefit of fog computing is that it reduces network traffic and provides platform for filtering and data analytics. Fog computing provides resources and services at the edge of the network or maybe even at end devices. This improves the Quality of Service and reduces service latency, therefore resulting in enhanced overall user-experience. The distinctive characteristics of fog computing are Location awareness and low latency, Geo-distributed, Mobility support, Real-time data analytics, Interoperability and Heterogeneity.

The importance of fog computing is that it processes data locally before transmitting to the cloud. IoT data are processed in smart devices or a data hub closer to the device which generating it. In cloud, accessing data always requires bandwidth allocation and we had to depend upon cloud repository, whereas in Fog computing, we can access data locally between devices independent on the cloud repository. Hence accessibility will be improved.

Fog computing allows the edge node devices to carry out the Local data processing, Cache data management, Dense geographical distribution, Local resource pooling, Load balancing, local device management, latency reduction for better QoS and edge node analytics. The critical aspect considered in the working of fog computing is Time. The most time-critical data are locally and it results in reduced latency and prevents from occurrence of major damages (E.g. Alarm status, Device status, Fault warnings). The less time-critical data are sent to the central mainframe which results in persistent, periodical storage that can be retrieved as and when required (E.g. Files, Reports for historical analysis, Device logs). The key advantage of using Fog computing in vehicular monitorization is that it offers many advantages over cloud computing, while the core technology is as such in cloud.

Figure 2 shows the infrastructure proposed in this paper. The proposed infrastructure presents a way for RSU to transform travel related information to fog instead of cloud where it can utilize the whole advantage of fog so far discussed. Generally, the transformation of information from RSU to centralized server may require some enumerations or preprocessing of data, which can be easily and optimally carried out in Fog layer. Fog will overcome the limitations of cloud, not as a replacement of it. By this way, instead of sending anything and everything to cloud, only data needed for long term analysis can be sent.

4 Results and Discussions

4.1 RSU Utilization After ARP

The ARP algorithm is proposed in a generic way for provisioning the resources, which in this paper is applied to the vehicular monitorization problem treating the RSUs as resources. Therefore, after applying ARP to the vehicular monitorization, the optimal utilization of RSUs is shown in the table. We have chosen iFogSim for simulating the proposed algorithm [12]. RSUs are created as fog nodes in the

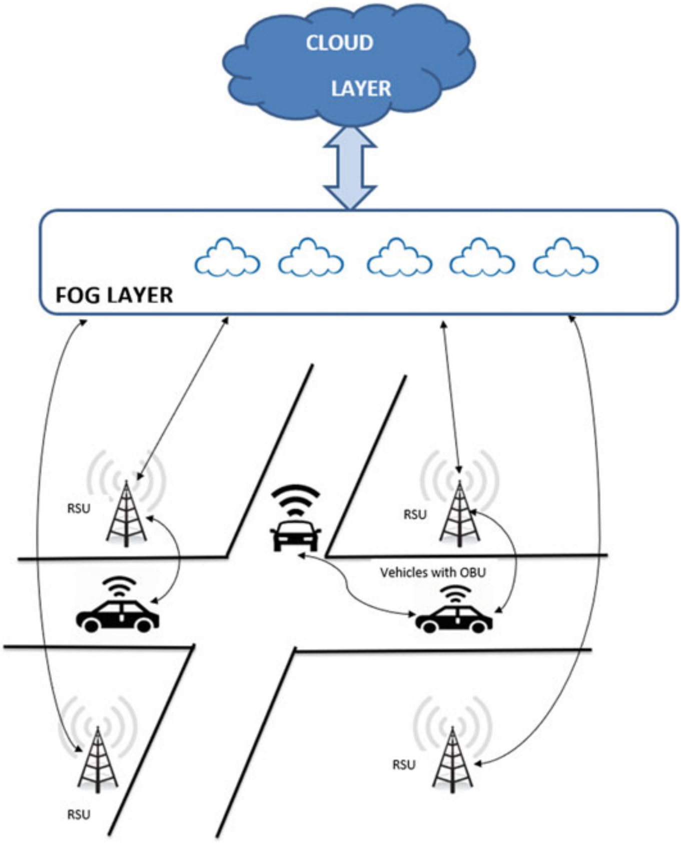


Fig. 2 Proposed infrastructure for vehicle monitorization

Table 1 RSU utilization using ARP (for $m = 10$ lanes)

Lane number	1	2	3	4	5	6	7	8	9	10
Binary code	0	1	0	1	0	0	1	0	1	0
Solution after ARP	−1	1	−1	0	−1	−1	2	−1	−1	0

iFogSim simulation environment and the number of lanes is also predefined in this approach.

Table 1 shows the optimized number of RSU utilized by using ARP algorithm for 10 lanes. The binary code ‘1’ indicates the presence of RSU in that lane and ‘0’ indicates no RSU for that lane. The ‘−1’ in the solution indicates that we do not need RSU for that lane. ‘0’ indicates that the RSU will cover that lane alone, ‘1’ indicates that the RSU in that lane cover up one adjacent lane on both sides and ‘2’ indicates that RSU in that lane cover up two adjacent lanes on both sides.

4.2 Performance Analysis

In our proposed methodology all the process is carried out on the fog environment. All the mentioned parameters are analysed and obtained in the iFogSim Simulator. For analysing the performance results, we run the approach for multiple times and it is visualized. The necessary information for the vehicle travellers are transmitted to them without any interruption. Since RSU is the essential part in transformation of vehicular information, the optimal location for placing the RSU is an important factor to be found. Different values for multiple scenarios are generated using simulation for a given lanes. The proposed system is analysed by adopting ARP algorithm in iFogSim simulator.

In the analysis, parameters such as μ, λ, k is treated equally. Figure 3 shows that number of RSU's utilized is low when optimized using ARP algorithm. It also ensures an effective communication and transformations of data. The values for RSU without ARP are obtained by considering the binary code for all lanes as 1. When the number of lanes is less (say 10 or 20), the ARP does not show any improvements, whereas when the number of lanes increases, the ARP shows optimal results by utilizing very less RSUs. This proves the solution is good when covering more number of lanes. Figure 4 shows that the power consumed by the RSU is less for optimized number of RSU using ARP when compared with deployment of all RSU's. Similarly, Fig. 5 shows that cost of installation of RSU is lower when we minimized the number using ARP. Figures 3, 4 and 5 show the effectiveness of ARP algorithm in providing an optimal number of RSU when the number of lanes is increasing rapidly. It is also clearly seen that the proposed ARP algorithm achieves a considerable efficiency in utilizing the RSU such that it has a maximum coverage while significantly reducing the installation cost and power consumed by RSU's.

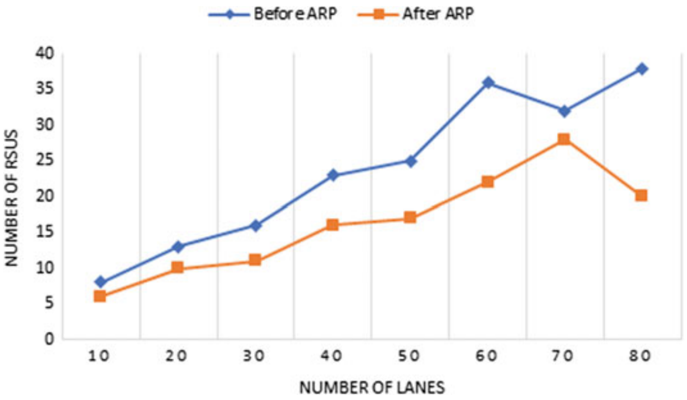


Fig. 3 Efficiency of ARP in RSU utilization

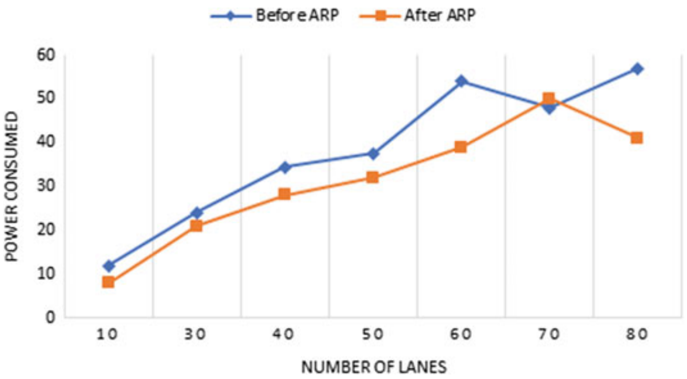


Fig. 4 Efficiency of ARP with respect to Power consumption

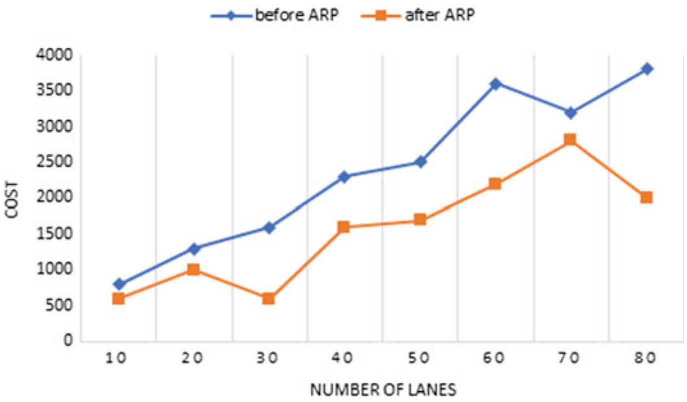


Fig. 5 Efficiency of ARP with respect to cost

5 Conclusion

This paper optimally minimizes the number of RSUs utilized in the vehicular monitorization application by using an efficient ARP approach, ensuring that, it covers all the lanes. This approach also finds the right locations to place the RSUs to cover all the lanes. And for the processing of the traffic and vehicular data, a superimposed fog computing layer is suggested and an infrastructure is also proposed. RSU are created as fog nodes in the simulation environment and the simulation results shows that the number of RSUs has been minimized considerably after applying ARP algorithm and the installation cost as well as the power consumption of RSU is minimized. The usage of fog computing layer for processing of data ensures low-latency services. The proposed approach can be expanded by using intelligent techniques such as PSO and other search techniques for handling different type of IoT data and applications that are deployed in Fog computing layer.

References

1. Kim, D., Velasco, Y., Wang, W., Uma, R.N., Hussain, R., Lee, S.: A new comprehensive RSU installation strategy for cost-efficient VANET deployment. IEEE Trans. Veh. Technol. **66**(5), 4200–4211 (2017)
2. Kai, K., Cong, W., Tao, L.: Fog computing for vehicular ad-hoc networks: paradigms, scenarios, and issues. J. China Univ. Posts Telecommun. **23**(2), 56–65 (2016)
3. Stojmenovic, I., Wen, S.: The fog computing paradigm: scenarios and security issues. In: Computer Science and Information Systems (FedCSIS), Federated Conference, pp. 1–8 (2014)
4. Arkian, H.R., Diyant, A., Pourkhalili, A.: MIST: fog-based data analytics scheme with cost-efficient resource provisioning for IoT crowdsensing applications. J. Netw. Comput. Appl. **67**, 152–165 (2017)
5. Fog computing and the Internet of things: extend the cloud to where the things are. White Paper. San Jose, CA, USA: Cisco (2015)
6. Bonomi, F., Milito, R., Zhu J.: Fog computing and its role in the Internet of things. In: ACM Workshop on Mobile Cloud Computing (MCC'12). New York, USA, ACM (2012)
7. Sankaranarayanan, M., Mala, C.: Genetic algorithm based efficient RSU distribution to estimate travel time for vehicular users. In: IEEE International Conference On Soft Computing and Machine Intelligence. IEEE (2015)
8. Ali, G.M.N., Mollah, A.S., Samantha, S.K., Mahmud, S.: An efficient cooperative load balancing approach in RSU based vehicular ad hoc networks (VANETs). IEEE International Conference on Control System. Computing and Engineering, pp. 28–30. Penang, Malaysia (2014)
9. Kaur, R.: TDB based optimistic RSUs deployment in VANETs. J: Comput. Sci. Eng. Technol. **4**(6), 708–712 (2013)
10. Salvo, P., Cuomo, F., Baiocchi, A., Bragagnini, A.: Road side unit coverage extension for data dissemination in VANETs. In: IEEE Conference on Wireless on-demand Network Systems and Services (WONS). IEEE (2012)
11. Ali, G., Chan, E., Li, L.: On scheduling data access with cooperative load balancing in vehicular ad hoc networks (vanets). J. Supercomputing **67**(2), 438–468 (2014)
12. Gupta, H., Dastjerdi, A.V., Ghosh, S.K., Buyya, R.: iFogSim: A Toolkit for Modelling and Simulation of Resource Management Techniques in the Internet of Things, Edge and Fog Computing Environments. Wiley (2017)

Identification of Meme Genre and Its Social Impact

Abhyudaya Pandey, Sandipan Mukherjee, Sandal Bajaj and Annapurna Jonnalagadda

Abstract Memes have taken the center stage as far as present day social media network is concerned. Every second, about a million memes are shared on various social media. In this paper we analyze a dataset which consist of data related to meme sharing and corresponding phrases used in the process. Popularity of memes is tracked followed by timestamps to identify the most shared memes across the internet. Meme sharing or propagation is also based on some hidden patterns like a social incident, a soccer game, a war or speeches by a public figures. The paper aims to produce a multivariate graph with the source of memes as nodes which are connected to other sources sharing the same meme using edges. The graph is then analyzed to identify patterns and genre of popular memes.

1 Introduction

In present-day social media, memes have become a very important attribute of expression of thoughts. Every second, on an average, about a million new memes enter the domain of social media. Users of social media share these memes with other users in their posts or via pages or blogs which are present abundantly on the Internet. The sharing of memes help in analyzing their lifecycle and the popularity they gained through it. Sustenance of a given meme depends on the impact it makes on a community that it dwells in.

A. Pandey · S. Mukherjee (✉) · S. Bajaj · A. Jonnalagadda
School of Computer Science and Engineering, VIT, Vellore, Tamil Nadu 632014, India
e-mail: sandipan.mukherjee2015@vit.ac.in

A. Pandey
e-mail: abhyudaya.pandey2015@vit.ac.in

S. Bajaj
e-mail: sandal.bajaj2015@vit.ac.in

A. Jonnalagadda
e-mail: annapurna.j@vit.ac.in

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_2

Some social media users subtly register or ignore these thought conveying pictures and texts, while others take them very seriously [10]. This paper [10] is based on the statistical study of propagation of memes by different users of social media and identifies the phenomena that resulted in the popularity of memes.

The term meme (meem) was first used by biologist Bauckhage [1] in his book The Selfish Gene. Memes have been further described by him as phenomena that rapidly gain popularity and notoriety in the Internet. Often, it so happens that the modifications add to the profile of the original idea that develops it into a phenomenon that "transgresses" both social and cultural boundaries. In present day social media, memes have become a very important attribute of expression of thoughts. Depending on the popularity or sharing that the memes gain, they sustain and go viral, get removed or get obsolete.

Society and users of social media are affected by the memes that they view at various levels as mentioned in the article "A Study of meme propagation: Statistics, research, authorities and spread by Dalal, O., Mahajan, D., Segall, I. and Vishvanath, M". The trend to track memes [2] on the Web ranges from tracking topic shifts over large time scales to unexpected spikes in the appearance of memes. There have been significant developments in this domain in the last decade. Scientists have developed framework for tracking short and distinctive phrases and have developed clustering algorithms for clustering textual variants of phrases. Over the years, the tracking of the societal phenomena [3] has been analyzed using media-related texts. The analysis reveals the underlying generative mechanism. Media artifacts have been assumed to reflect environmental dynamics. It has been observed [4] that the prevalence of social media has bridged the "dissemination and proliferation" of memes. Hence scientists and researchers have developed tools to rank memes based on their popularity that not only helps to reduce data irregularity but also act as an aid for online advertising. The advent of tracking propagation of memes has been an important topic of research since 1986, when Richard Dawkins, "had caught the spirit of the new age." [1, 5] According to previous studies [5], meme popularity distribution has been described by "heavy tailed distribution" or "power law". Researchers have criticality studied growing and non-growing networks based on the completion that has been induced by the criticality model. As mentioned before, tracking memes in the mass social media has been a developing trend in the recent times. There have been researches to conclude whether the fact: "false rumor propagates through Twitter and the truth propagates between friends in Facebook" is true or not. To conclude, data scientists have used the concept of competing memes in composite networks. In such analysis [6], it is assumed that each meme exists in a SIS-like propagation model and to study such a system, one computes the nonlinear dynamic system (NLDS). In this, a metric is developed for each meme that is based on the eigen value derived from the respective matrix and this matrix is used to determine the "winning" matrix. Further, there has been advent of websites that help us in studying meme propagation in large-scale social media using streams of micro-blogging data. However, such websites are not designed for the scientific community and are destined for end users. In recent times, there have been developments of a unified framework [7] to bridge this gap between the scientific community and the interface destined for end users. In a nutshell, this framework model streams social network data as a series of events in which the users and the memes are the participants. Such a model will facilitate comparison of high level of statistical features across different communities in Web 2.0. While the primary approach has been tracking the memes, some researchers [8] have gone as far as regarding memes as “potential troublemakers”. Three communication-oriented metrics have been addressed to answer the problem; they are content, form and stance. From this metrics one can derive the possible paths for further meme-oriented analysis from the digital world. However, as mentioned before, the primary aim has always been to use the existing link structure to track the flow of data available in social media and determine potential threat to the community or the society, regarding the phenomenon as an “infection”. The process of tracking data flow and analyzing the same in social media [9], however, is not only concentrated on memes but also the phrases attached to the memes. It spans right from someone commenting on someone’s post to someone mentioning about a fact in his or her blog or post. Thus, one need not necessarily share a meme in social media, but can be regarded as its source if he or she mentions about it in his or her blog. The beauty of analyzing social media by tracking dataflow lies in this statement. Despite not having the intention of infecting the social media, one may become the victim and he or she can be tracked efficiently.

The authors of the papers discussed above had a mathematical approach in analyzing the meme-tracking data. Various methods and approaches to analyze the data have been discussed. While some have discussed why it is necessary to track memes, some have inferred upon a “winning” meme in the social media domain. As far as the frameworks to deal with memes are concerned, they have been building for the benefit of the scientific community and end users of the memes. The fact that memes can be a potential threat to the society can be inferred from some papers as well. But, a detailed study of the reason of their popularity, background that caused the meme to be propagated and the impact that the meme had has never been discussed. This is what this paper aims to discuss.

We use the Meme-tracking system to understand how phrases spread across news sources. We develop an algorithm and corresponding modules, to observe

1. the time at which the memes have been shared
2. the memes occurring most number of times
3. number of times a meme has been shared or has been in use
4. correlate manually the social cause (if any) and the social impact of the meme sharing
5. build a multivariate graph with sources as nodes and the link between the sources as edges
6. using this multivariate graph, determine communities and manually interpret genre
7. visualizing the data flow using these graphs

and observe how the category of the meme affects its life. We begin by looking at various statistics from the frequency data(for points 1, 2, 3, and 4) to understand differences between the lifetime of memes across subject matter and media source, and look for methods to predict how the shape of the graph will change over time. [10] Then, using the time data we consider the graph of news sources and directed edges from the sources that are hyperlinked in the original article. Additionally, we built a theoretical example to determine the actual influence network. *We assume that this composite network is a SIS like propagation model* [6]. We approach the system in a statistical and not in a direct mathematical way. Thus, this paper is based on the statistical study of propagation of memes by different users of social media and identifies the phenomena that resulted in the popularity of memes and to identify patterns, if any. We discuss the approach and the methodology to obtain the above mentioned results in the following sections.

2 Methodology

Let $D_1 = (A, M, T)$ represent a dataset where A represents article id assigned to a given meme M shared at time T. Let $D_2 = (A, P)$ represent a dataset in which A represents article id and P represents phrases. The notion being, for a given meme M_1, shared at time T_1 an article id is assigned. Let P_1 be the phrase used while sharing M_1 in social media. Let M_2 be another meme shared at time T_2. Let P_2 be the phrase used while sharing the meme. If P_1 and P_2 are same, then P_2 is assigned the same article id as P_1. Using these datasets D_1 and D_2, **frequency data** is computed that comprises of statistical results such as number of times a meme has occurred, identification of the most popular meme and the timestamps in which it has occurred. Further, D_1 and D_2 is used to compute **time data** which is then used to simulate a multivariate graph $G_1 = (S, E)$, where meme sources S are the nodes and E represents the edges between the meme sources. An edge is created between a pair of meme sources Sif and only if a common phrase has been used while sharing these memes. Another graph, $G_2 = (v, e)$ is constructed, where v is a meme source and e is an edge which is constructed if a particular meme has its source or inspiration from another meme (Fig. 1).

2.1 Frequency Data

First goal is to identify the magnitude of posts containing a given meme and how it changes over time and analyze the statistics that will simplify understanding of the nature of the media cycle.

The basic interest is to obtain the differences in the key features of the meme propagation and thereby determine the social impact of the “winning” meme based on the statistical analysis.

2.2 Time Data

A multi-graph is then generated on a small subset of phrase cluster data. In this graph, each node refers to a source (article_id) and two sources using the same root phrase share an edge which is labeled with the phrase. Although represented as a graph, the structure simplifies into a set of levels, where the nth-level connects entirely with every node in levels 1 to n (this hereby forms a complete graph on every node that uses the phrase). Second goal lies in understanding the influence of the media-blog network.

2.3 Modular Design

See Fig. 1.

2.4 Proposed Algorithm

In this section, we propose the algorithms in order to find the frequency data and the time data.

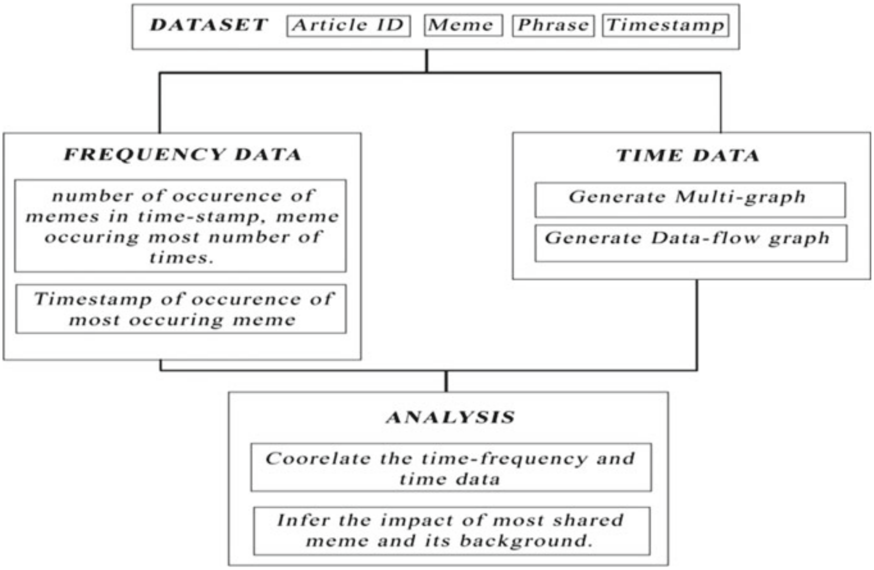

DATASET
Article ID
Meme
Phrase
Timestamp
FREQUENCY DATA
number of occurence of memes in time-stamp, meme occuring most number of times.
Timestamp of occurence of most occuring meme
TIME DATA
Generate Multi-graph
Generate Data-flow graph
ANALYSIS
Coorelate the time-frequency and time data
Infer the impact of most shared meme and its background.

Fig. 1 Block model of meme analysis

2.5 Algorithm SAS

The algorithm accepts the MemeTracker data as its input in the form of a SQLite database. As mentioned in the modular design, this algorithm comprises of two modules, the frequency data module and the time data module. Both frequency and time data module accepts this MemeTracker data as their input. The frequency data module uses this dataset that computes the number of times a meme has occurred in the dataset by extracting the article id from the table named quotes in the MemeTracker data and storing them in a temporary array, say array1. Also, article id is extracted from the table named articles in the same dataset and stored in a temporary array, say array2. Then the elements of both the arrays are compared and if the article ids match, then the counter keeping track of the occurrences of that particular meme, is incremented. This counter is stored in another array, say array3. In array3, the ith element corresponds to the ith element of array2. Thus, the number of times the meme has occurred is calculated. This result is then plotted in the form of a graph. Then, using the data available in array3, the meme that has occurred most number of times is calculated. Then, the timestamps at which this meme (the meme that has occurred most number of times) has been shared is plotted in a graph by extracting the dates column from the articles table corresponding to the article id of the meme. If two memes have occurred equal number of times, then the meme that occurs first in the dataset is considered.

Similarly, in the time data module, article id and phrases are extracted from the table quotes. Initially, the meme sources are plotted as nodes and named corresponding to its article id. Then, comparing the phrases corresponding an article id from the table quotes, the edges are constructed if a pair of nodes (article id) has the same phrase. This is followed by construction of another graph, known as the dataflow graph. This is constructed using the links table, by extracting the link in column from the links table. This column (link_in) consists of article id of the meme related to the present meme being read in the dataset.

Let us consider the graph of time data (occurrence v/s time). From this graph, we are able to infer which meme was shared at what time stamp. Let us consider another graph, the time frequency graph (no. of occurrences vs timestamp) of a particular meme. Super imposing the first graph on the second or vice versa, we can infer when a particular meme was most shared. Then, we manually co-relate the events around this time that are related to the propagation of the meme. Thus, we infer the impact and the cause of propagation of the meme by super imposing the time data graph on the time frequency graph and vice versa.

We discuss about the results obtained using the SAS algorithm mentioned above in the following section.

3 Results and Discussions

In this section, we analyze the results obtained using the SAS algorithm on the MemeTracker data. The MemeTracker data tracks memes over various news media and blog sources during the timeline of August 2008 till April 2009. The SAS algorithm discussed in this paper helps in analyzing the data flow, which is very important to interpret impact of memes in the society. The implementation has 3 stages: data collection, data correlation, and prediction.

Stage 1: The existing memes in social media and blogs are stored in a database along with the corresponding phrase used while posting it. This storage is done in such a way that each different meme has a unique identification number attached to it. Each unique phrase has a unique id attached to it. It also has another attribute that connects it to the meme.

Stage 2: Data correlation. Using the data available, we correlate the time data and time frequency values. We also use the data available to infer the genre the meme belongs to and also the impact it has.

Stage 3: Finally, using artificial intelligence and machine learning techniques, we use the existing results available to predict the impact of a meme that is about to be uploaded in social media. The novelty lies in the fact that using this algorithm, we can prevent social issues happening because of the meme.

For example, when a political leader comments about the opposition, it might result in creation of a number of memes that might be shared on social media. The impact of this can be positive or negative. The result of this analysis is as discussed in the Fig. 2.

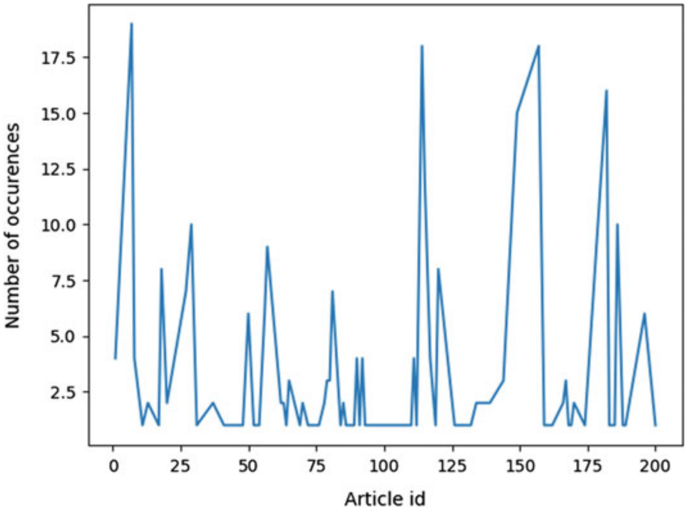


Fig. 2 Given plot depicts the number of times a meme was shared on different platforms and blogs (time frequency)

Here, the x-axis represents the article id assigned to memes and the y-axis represents the number of times a meme was shared during its entire lifecycle. One can clearly visualize the difference in popularity of memes by studying this graph.

Using a threshold value for *y*-axis, the popular memes can easily be identified as well.

The lifecycle of a meme is inferred using the algorithms available in the paper and is as displayed in Figs. 3 and 4.

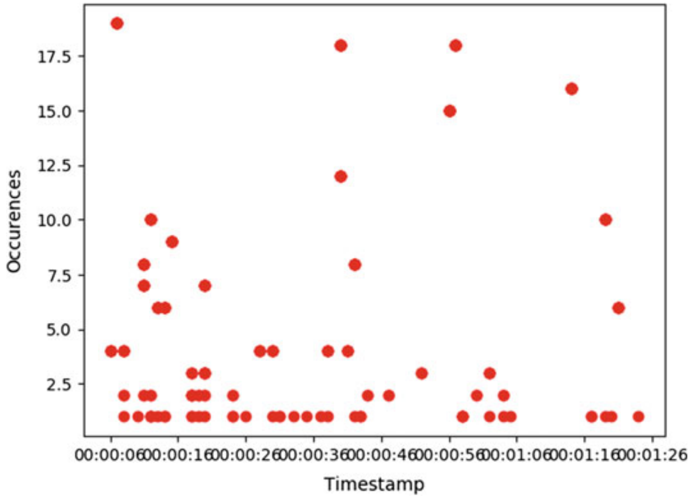


Fig. 3 Given plot shows the time at which a meme was shared (time frequency)

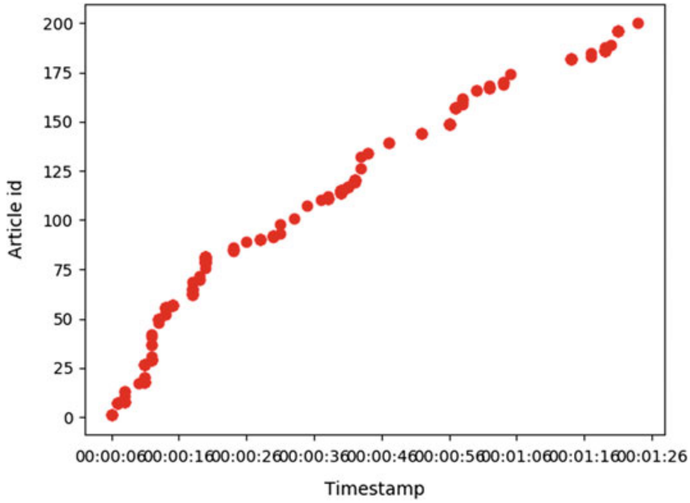


Fig. 4 Plot shows the time at which memes were shared according to their article ids (time data)

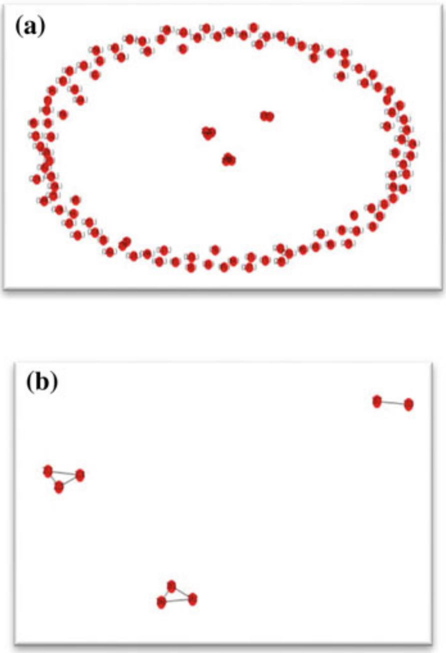


Fig. 5 Meme tracking

In Fig. 3, the *x*-axis denotes time and *y*-axis denotes number of times the meme was shared. This graph is constructed for individual memes. Correlating the time and the number of times a particular meme was shared during that time, one can correlate it with incidents in the background that resulted in the popularity of the meme.

In Fig. 4, the *y*-axis represents article ids assigned to memes and the x-axis represents the **first** timestamp in which it was shared. One can infer upon the lifecycle of a meme using this graph.

Further, when a group of social media users or bloggers share this meme using a specific hash tag or a phrase, this information is tracked using the algorithms mentioned above (Fig. 5a, b). Thus, a community of meme sharers identified.

Figure 5a, b represents multi-graphs that depict the various genre of memes present in the social media domain.

The genre of the most popular meme is easily identified by analyzing the phrase it is shared with which can vary over politics, sports, economics, etc. The reason of creation of a meme is inferred from its source and by tracking its flow in the social environment (Fig. 6).

Figure 6 represents the dataflow graph, where meme sources act as nodes and the fact that same phrase was used to share them as edges. Using the dataflow graph, a relation, if any, among the reason the meme propagated can be identified.

Figure 7a represents the multi-graph constructed by treating memes as nodes. A pair of node is connected if they are shared using the same phrase. A number of communities can be seen in the multi-graph generated. Comparing the phrases

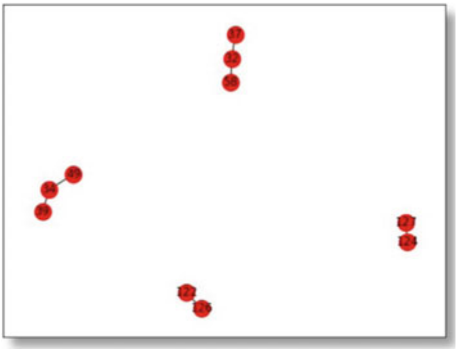


Fig. 6 Magnified view of data flow between sources

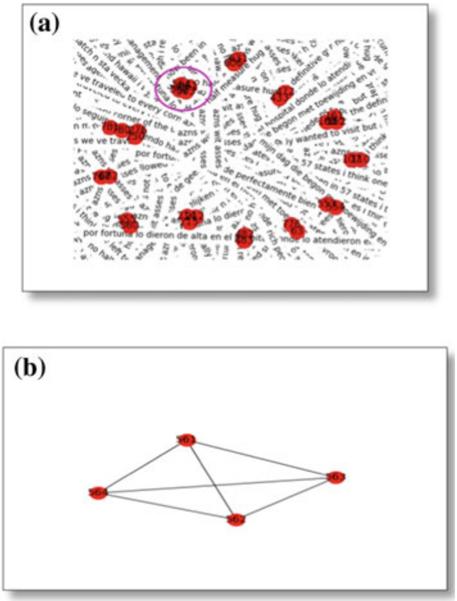


Fig. 7 **a** Multi graph representing meme sources as nodes and phrases connecting them as edges, **b** magnified version of encircled region in Fig. 7a

that join a pair of nodes in the communities in the multi graph, a specific genre can be identified. Figure 7b represents one such genre, a political genre, which was identified manually. This result can be further extended and can be correlated with the incidents that occurred around the same time and hence its popularity can be analyzed and visualized, as in Fig. 3.

However, this manual analysis can be computed using data analytics software and tools and the result can be inferred using artificially intelligent algorithms. The amount of data available is a bottleneck. Resources required for analyzing such large amount of data and the corresponding algorithms used while decomposing the dataset is also a topic of research. This and other scopes of developments are discussed in the following section.

4 Conclusion and Future Prospects

In this paper, we analyzed the dataset of memes from MemeTracker and obtained two subsets of data. These two datasets, Time data and Frequency data are then correlated understand the relation between the time and the popularity of propagating memes on social media. We then segregated the dataset manually into different genres and we indentified the social cause and impact behind the meme design.

A number of parallel and distributed algorithms can be developed to deal with the problem statement, considering the huge amount of data available in the dataset that needs to be analyzed. As mentioned earlier, the genre identification of memes and event recognition determining the inspiration behind it was manually concluded. This can be done by implementing Artificial Intelligence and training the system to identify the same. It will be wise to say that this paper just deals with the tip of the iceberg. In order to extract, analyze, model and interpret the entire dataset, one needs multi-disciplinary knowledge in various fields of computer science and mathematics. Thus, with more in-depth research, a system can be looked forward to which would predict the exact effect and lifecycle of a meme or phrase even before it enters into the world of social media system.

References

1. Bauckhage, C.: Insights into internet memes. In: ICWSM (2011)
2. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM (2009)
3. Dooley, K., Corman, S.: Dynamic analysis of news streams: institutional versus environmental effects. Nonlinear Dyn. Psychol. Life Sci. **8**(3), 403–428 (2004)
4. Gleick, J.: What defines a meme? Smithson. Mag. (2011)
5. Kim, Y., Park, S., Yook, S.-H.: The origin of the criticality in meme popularity distribution on complex networks. Sci. Rep. **6**, 23484 (2016)
6. Wei, X., et al.: Competing memes propagation on networks: a case study of composite networks. ACM SIGCOMM Comput. Commun. Rev. **42.5**, 5–12 (2012)
7. Kwak, H., et al.: What is twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web. ACM (2010)
8. Shifman, L.: Memes in a digital world: reconciling with a conceptual troublemaker. J. Comput.-Mediat. Commun. **18**(3), 362–377 (2013)
9. Adar, E., Adamic, L.A.: Tracking information epidemics in blogspace. Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence. IEEE Computer Society (2005)
10. Dalal, O., et al.: A study of meme propagation: statistics, rates, authorities, and spread

Shannon's Entropy and Watershed Algorithm Based Technique to Inspect Ischemic Stroke Wound

V. Rajinikanth, K. Palani Thanaraj, Suresh Chandra Satapathy, Steven Lawrence Fernandes and Nilanjan Dey

Abstract Ischemic Stroke (IS) is usually initiated due to the neurological shortfall in human brain and which can be recognized by inspecting the periphery of the brain sections. In this paper, a two-step procedure is proposed to extract and evaluate the IS injury from brain Magnetic Resonance Image (MRI). In the initial step, Social Group Optimization and Shannon's entropy based tri-level thresholding is executed to enhance the IS section of the test image. During the second step, enhanced IS section is then mined using the marker-controlled Watershed (WS) algorithm. The proposed practice is tested on benchmark ISLES 2015 dataset. Performance of the WS segmentation is also verified with the segmentation approaches, like seed-based region growing (SRG) and the Markov Random Field (MRF). The outcome of this study authenticates that, WS provides enhanced picture likeness indices, like Jaccard (90.34%), Dice (94.92%), FPR (7.77%) and FNR (2.65%) compared with SRG and MRF.

V. Rajinikanth (✉) · K. Palani Thanaraj
Department of Electronics and Instrumentation, St. Joseph's College
of Engineering, Chennai 600119, Tamil Nadu, India
e-mail: rajinikanthv@stjosephs.ac.in

S. C. Satapathy
School of Computer Engineering,
Kalinga Institute of Industrial Technology (Deemed to be University),
Bhubaneswar 751024, Odisha, India

S. L. Fernandes
Department of Electronics & Communication Engineering, Sahyadri College
of Engineering & Management, Mangalore, India

N. Dey
Department of Information Technology, Techno India College
of Technology, Kolkata 700156, India

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_3

1 Introduction

In recent years, a number of automated medical image segmentation and classification approaches are proposed and implemented by the researchers [1, 2]. These approaches will assist the diagnosing clinics and doctors to obtain the vital information about the disease-infected region and its infection rate. The premature discovery and analysis of various diseases also helps to reduce the morbidity and mortality rates. Due to these reasons, computer-assisted disease examination procedures are widely considered by the researchers. In this paper, Ischemic Stroke (IS) injury examination is considered for the investigation.

Generally, stroke is caused due to neurological deficit and interruption of blood delivery because of fault in brain blood vessels [3, 4]. To arrange better monitoring and therapeutic procedure, it is necessary to evaluate the cause and position of abnormality. If correct region of irregularity is identified, the physician can implement feasible treatment to cure the sick person [5].

In the literature, a number of approaches are proposed to mine noteworthy information related to IS injury. Usinskas and Gleizniene [6] implemented a six-step computer-based assessment practice to sense the center of IS injury with 2D CT brain picture. Kabir et al. implemented a procedure using T2, Flair, and DW pictures to extract stroke region with EM-MRF approach [7]. Rajinikanth and Satapathy proposed a comprehensive study with various image extraction schemes [8]. Yahiaoui and Bessaid discussed fuzzy-C-means technique to extract abnormality from CT pictures [9]. Mitra et al. executed Bayesian-MRF to mine stroke wound form MRI and lastly isolate the pictures using a classifier [10]. The work by Kanchana and Menaka presented automated IS injury detection from CT and MRI [11]. Research of Maier et al. implements various classifiers to separate Ischemic Stroke Lesion Segmentation (ISLES 2015) challenge images into different groups [12, 13]. Earlier works also confirms the execution of IS from previous study, CT/MR pictures with various approaches [14].

This paper proposes a tool to mine and evaluate stroke section from MRI dataset. Initially, enhancement of stroke province is done with Shannon's entropy based tri-level thresholding. In this work, the recently proposed Social Group Optimization (SGO) practice is implemented to get optimal thresholds of test picture by maximizing the Shannon's entropy function. Later, the Watershed (WS) algorithm based segmentation is considered to mine the infected part in thresholded test image [15]. The efficiency of WS methodology is validated against the seed based region growing (SRG) technique [16] and the Markov Random Field (MRF) approach [1].

2 Methodology

SGO motivated practice to segment the stroke section from 2D MRI picture is executed in this paper. The picture processing task is implemented with the combination of thresholding and segmentation practices.

2.1 *Social Group Optimization and Shannon's Entropy*

Satapathy and Naik developed SGO in 2016 to find optimal solution for benchmark problems [17]. SGO is based on the replica of human group activities and it contains following sections; (i) Civilizing stage and (ii) Knowledge attaining.

The arithmetical appearance of SGO is presented below with Eqs. (1)–(5):

Let

$$G_{\text{finest}} = \text{maximum}\{f(H_v)\text{for } v = 1, 2, \ldots N\}, \tag{1}$$

where H_v denotes the preliminary data the people have, v =1, 2, 3,…, N specify entire strength of collection, and f_w denotes cost function.

The civilizing stage will alter the orientation of the agents as given below;

$$H_{\text{updated}_{v,w}} = c * H_{\text{initial}_{v,w}} + R * (G_{\text{finest}_w} - X_{\text{initial}_{v,w}}) \tag{2}$$

where H_{updated} represents renewed location, H_{initial} is the early location, G_{finest} specifies global position, R is random figure [0, 1] and c is self-introspection value with a choice [0, 1]. The constant c assigned with 0.2 [18].

At the second stage, citizens are motivated to achieve the global position as follows;

$$H_{\text{updated}_{v,w}} = X_{\text{initial}_{v,w}} + r1 * (H_{v,w} - H_{R,w}) + r2 * (G_{\text{finest}_w} - H_{v,w}), \tag{3}$$

where $r1$ and $r2$ represents arbitrary values of [0, 1] and $H_{R,w}$ is randomly assigned position a citizen. During the experimentation task, SGO finds the optimal threshold by maximizing the Shannon's function [19, 20].

Multi-thresholding groups matching image pixels according to requirement [21]. Tri-level thresholding improves the visibility of tumor existing in MRI [1, 21]. Later, a chosen segmentation practice can be implemented to extract the tumor. This research adopts the Shannon's function proposed in [20] to preprocess the brain MRI.

Shannon's function can be expressed as follows;

Consider a picture of dimension $M * N$. The gray-level pixel organization (X, Y) is formulated as $G(E, F)$, with $E \in \{1, 2, \ldots, m\}$ and $F \in \{1, 2, \ldots, n\}$. If K represents total gray values of the picture and the position of all gray pixels $\{0, 1, 2, \ldots K - 1\}$ can be represented as O, as:

$$G(E, F) \in O\ \forall (E, F) \in \text{image} \tag{4}$$

Then, the regularized histogram is; $S = \{s_0, s_1, \ldots s_{K-1}\}$.

For tri-level thresholding, it can be framed as;

$$S(T) = s_0(t_1) + s_1(t_2) + s_2(t_3), \tag{5}$$

where, $T = \{t_1, t_2, \ldots, t_K\}$ is the threshold value and $T*$ the final threshold.

2.2 Segmentation Approaches

Marker-driven Watershed (WS) approach presented by Roerdink and Meijster is adopted to extract the stroke region [15]. WS is grouping of Sobel border discovery procedure, marker driven morphological procedure and mining. The clear explanation of the adopted WS can be found in [22].

The outcome of the watershed algorithm is then validated against the alternative approaches, like the seed based region growing method [16] and the Otsu's based Markov random field segmentation procedures existing in the imaging literature.

2.3 Skull Stripping

SS is used to eliminate the outer section of brain, before other processing is implemented. The SS assist in removal of skull part with the help of a threshold-filter based by incorporating a mask, which protects the low intensity pixels [23].

2.4 Stroke Analysis

This paper implements SGO-based tool to examine stroke of MRI pictures. The superiority of proposed tool is confirmed with respect to the Ground Truth (GT). Initially, the picture resemblance measures, like Jaccard, Dice, FPR, and FNR are computed based on the following literature [21, 23].

The mathematical expression for similarities is shown in Eqs. (6)–(9);

$$\text{JSC}(\text{GT}, I) = \text{GT} \cap I / \text{GT} \cup I \tag{6}$$

$$\text{DSC}(\text{GT}, I) = 2(\text{GT} \cap I)/|\text{GT}| \cup |I| \tag{7}$$

$$\text{FPR}(\text{GT}, I) = \big(\text{GT} \big/ I\big)/(\text{GT} \cup I) \tag{8}$$

$$\text{FNR}(\text{GT}, I) = \big(I \big/ \text{GT}\big)/(\text{GT} \cup I), \tag{9}$$

where, I symbolize mined region.

Further, the image parameters, like sensitivity, specificity, accuracy, precision, Balanced Classification Rate (BCR), and Balanced Error Rate (BER) are also computed Eqs. (10)–(15) [24, 25]:

$$\text{Sensitvity} = T_{\text{P}}/(T_{\text{P}} + F_{\text{N}}) \tag{10}$$

$$\text{Specificity} = T_{\text{N}}/(T_{\text{N}} + F_{\text{P}}) \tag{11}$$

$$\text{Accuracy} = (T_{\text{P}} + T_{\text{N}})/(T_{\text{P}} + T_{\text{N}} + F_{\text{P}} + F_{\text{N}}) \tag{12}$$

$$\text{Precision} = T_{\text{P}}/(T_{\text{P}} + F_{\text{P}}) \tag{13}$$

$$\text{BCR} = 1/2\,(T_{\text{P}}/(T_{\text{P}} + F_{\text{N}}) + T_{\text{N}}/(T_{\text{N}} + F_{\text{P}})) \tag{14}$$

$$\text{BER} = 1 - \text{BCR}, \tag{15}$$

where, I_{GT} is GT, I_S is mined region, T_{N}, T_{P}, F_{N} and F_{P} signifies true-negative, true-positive, false-negative and false-positive; correspondingly.

3 Results and Discussion

This section illustrates experimental outcomes of proposed tool. In this study, two different stroke injury dataset is considered for the analysis. The first one is from the Radiopaedia image recorded with the well known modalities, such as T1, T2 and Flair [26]. This test image has a pixel size of 630 × 630. The second dataset is from the well-known ISLES 2015 challenge dataset with a size 77 × 77 pixels recorded with flair and diffused weighted modality [27]. In order to have a good visibility, these test images are resized into 256 × 256 sized images before the evaluation process. This image is also provided with GT. During the experimental investigation, flair modality based images are considered.

Initially, this tool is tested on the Radiopaedia dataset recorded with T1, T2 and Flair modalities. These images are associated with the skull section and in order to have better assessment, skull stripping is carried before the preprocessing operation. Figure 1 shows the results obtained with the T1 sample image. Initially the preprocessing is implemented based on the SGO + Shannon's tri-level thresholding. Next stage involves in extraction of the infected section from test picture. The post processing is initially executed with watershed approach as discussed in Sect. 2.2. Later, seed based region growing (SRG) and the Otsu's + Markov Random Field (MRF) approaches are implemented to mine the SI and the outcomes are clearly shown in Fig. 1h, j and l respectively. Alike practice is executed with other images and the outcome is presented in Fig. 2. The result confirms that observed that, WS and SRG approaches are competent in mining stroke section of MRI image, irrespective of its modality as shown in Fig. 2d, e. But, the MRF approach fails to extract the required information from T2 modality based MRI as depicted in Fig. 2f. This experimental approach also confirms that, CPU instance of WS approach is around

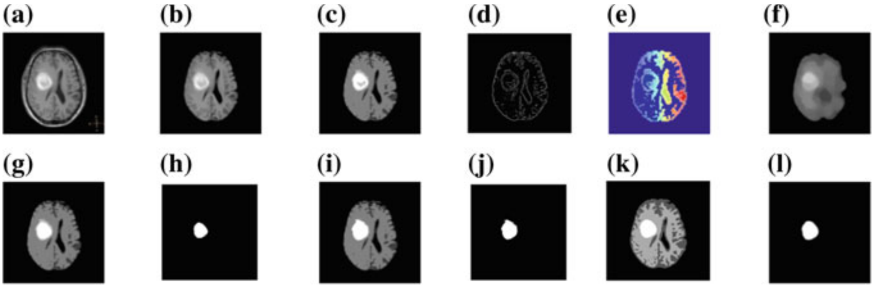


Fig. 1 Outcome of executed approach on the chosen test image. **a** Test picture, **b** skull eliminated image, **c** SGO + FE thresholded image, **d** sobel edge detection, **e** RGB color space with water shed, **f** morphological operation, **g** outcome of watershed segmentation, **h** tumor region by WS, **i** outcome of seed region growing, **j** tumor region by SRG, **k** outcome of MRF segmentation, and **l** tumor region by MRF

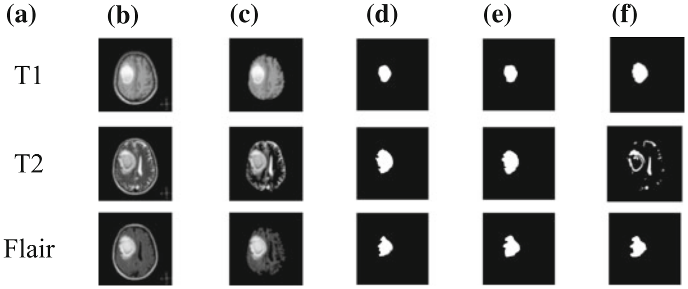


Fig. 2 Outcome attained using various modality MRIs. **a** Modality, **b** test image, **c** skull stripped image, **d** outcome of WS, **e** outcome of SRG, and **f** outcome of MRF

58.61 s and for the SRG, it is around 68.02 s respectively (computed using Matlab's Tic-Toc). The MRF approach shows an average CPU time of 71.57 s, which confirms that, the WS approach is efficient than the SRG and MRF techniques.

Efficiency of this tool is verified with test images of ISLES 2015 dataset. It is a three-dimensional (3D) dataset, in which the sample image slices are initially extracted by using the ITK-SNAP tool [28]. Later, the original 77 × 77 pixel images are upscaled to 256 × 256 sized images. Alike practice is implemented for GT. Figure 3 depicts the results of the chosen slices of flair modality ISLES 2015 dataset image. Figure 3b, c shows test image and GT. Figure 3d–f presents results by WS, SRG, and MRF techniques.

To evaluate the superiority of considered technique, comparison between mined stroke section and the GT is executed and picture similarity and quality measures are estimated as in Tables 1 and 2, and Fig. 4. These tables confirm that, WS procedure is efficient in extracting the abnormal section of MR images. The results of Fig. 4 also confirm that, the average result obtained with the WS technique is better compared to the alternative approaches considered in this paper.

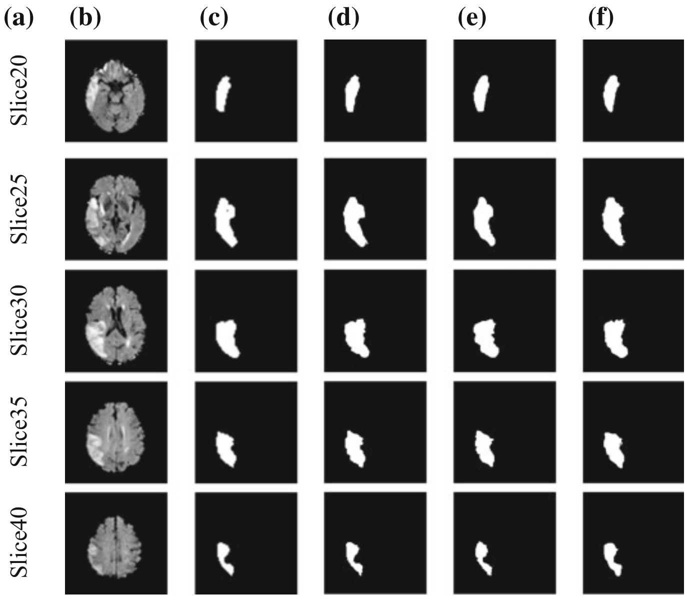


Fig. 3 Results attained with Flair MRI. **a** Slice number, **b** test image, **c** ground truth, **d** WS, **e** SRG, **f** MRF

Table 1 Image similarity measures obtained with flair modality

Image	Jaccard	Dice	FPR	FNR
Slice_{20}	0.8925	0.9432	0.0846	0.0320
Slice_{25}	0.9115	0.9537	0.0891	0.0073
Slice_{30}	0.9106	0.9532	0.0593	0.0355
Slice_{35}	0.9140	0.9551	0.0699	0.0221
Slice_{40}	0.8883	0.9409	0.0858	0.0354

Table 2 Image performance values of chosen ISLES 2015 pictures

Image	SEN	SPE	ACC	PRE	BCR	%BER
Slice_{20}	0.9972	0.9679	0.9825	0.9989	0.9825	1.7408
Slice_{25}	0.9940	0.9927	0.9934	0.9995	0.9933	0.6631
Slice_{30}	0.9964	0.9645	0.9804	0.9979	0.9805	1.9497
Slice_{35}	0.9969	0.9779	0.9873	0.9990	0.9873	1.2621
Slice_{40}	0.9978	0.9645	0.9811	0.9991	0.9812	1.8803

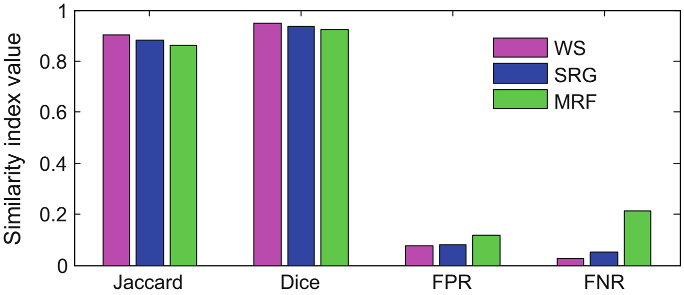


Fig. 4 Graphical representation of the average similarity index obtained with WS, SRG, MRF approaches

4 Conclusion

The work executes an analyzing tool by integrating Shannon's approach with Watershed (WS) extraction practice to enhance the performance of stroke injury examination. The ISLES dataset is adopted to evaluate the prominence of the implemented tool. The Radiopaedia MRI dataset is initially analyzed using the tri-level thresholding approach based on SGO. Later the proposed WS approach is validated against the SRG and MRF approaches. The results confirm that, the CPU time of WS is better compared with the SRG and MRF. An examination between extracted stroke section and GT offered better values of Jaccard (90.34%), Dice (94.92%), FPR (7.77%) and FNR (2,65%) compared with SRG and MRF. Hence, the proposed tool can be considered to examine the real-time clinical MRI images.

References

1. Palani, T.K., Parvathavarthini, B., Chitra, K.: Segmentation of brain regions by integrating meta heuristic multilevel threshold with Markov random field. Curr. Med. Imaging Rev. **12**(1), 4–12 (2016)
2. Rajinikanth, V., Raja, N.S.M., Satapathy, S.C.: Robust color image multi-thresholding using between-class variance and cuckoo search algorithm. Adv. Intel. Syst. Comput. **433**, 379–386 (2016)
3. Havaei, M., Davy, A., Warde-Farley, D., et al.: Brain tumor segmentation with deep neural networks. Med. Image Anal. **35**, 18–31 (2017). https://doi.org/10.1016/j.media.2016.05.004
4. Schmidt, M.A., Payne, G.S.: Radiotherapy planning using MRI. Phys. Med. Biol. **60**, R323–R361 (2015). https://doi.org/10.1088/0031-9155/60/22/r323
5. STROKE: http://www.world-stroke.org/
6. Usinskas, A., Gleizniene, R.: Ischemic stroke region recognition based on ray tracing. In. Proceedings of International Baltic Electronics Conference (2006). https://doi.org/10.1109/bec.2006.311103
7. Kabir, Y., Dojat, M., Scherrer, B., Forbes, F., Garbay, C.: Multimodal MRI segmentation of ischemic stroke lesions. In: 29th Annual International Conference of the IEEE Engineering in

Medicine and Biology Society EMBC. Lyon, France, (2007). https://doi.org/10.1109/iembs.2007.4352610

8. Rajinikanth, V., Satapathy, S.C.: Arab J Sci Eng (2018). https://doi.org/10.1007/s13369-017-3053-6
9. Yahiaoui, A.F.Z., Bessaid, Y.: Segmentation of ischemic stroke area from CT brain images. In: International Symposium on Signal, Image, Video and Communications (ISIVC) (2016). https://doi.org/10.1109/isivc.2016.7893954
10. Mitra, et al.: Lesion segmentation from multimodal MRI using random forest following ischemic stroke. NeuroImage **98**, 324–335 (2014)
11. Kanchana, R., Menaka, R.: Computer reinforced analysis for ischemic stroke recognition: a review. Indian J. Sci. Technol. **8**(35), 81006 (2015)
12. Maier, O., et al.: ISLES 2015—A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. Med. Image Anal. **35**, 250–269 (2017)
13. Maier, O., Schröder, C., Forkert, N.D., Martinetz, T., Handels, H.: Classifiers for ischemic stroke lesion segmentation: a comparison study. PLoS ONE **10**(12), e0145118 (2015)
14. Rajinikanth, V., Satapathy, S.C., Dey, N., Vijayarajan, R.: DWT-PCA image fusion technique to improve segmentation accuracy in brain tumor analysis, Lecture Notes in Electrical Engineering, vol. 471, pp. 453–462 (2018). https://doi.org/10.1007/978-981-10-7329-8_46
15. Roerdink, J.B.T.M., Meijster, A.: The watershed transform: definitions, algorithms and parallelization strategies. Fundam. Informaticae **41**, 187–228 (2001)
16. Shih, F.Y., Cheng, S.: Automatic seeded region growing for color image segmentation. Image Vis. Comput. **23**, 877–886 (2005)
17. Satapathy, S., Naik, A.: Social group optimization (SGO): a new population evolutionary optimization technique. Complex Intell. Syst. **2**(3), 173–203 (2016)
18. Naik, A., Satapathy, S.C., Ashour, A.S., Dey, N.: Social group optimization for global optimization of multimodal functions and data clustering problems. Neural Comput. Appl. (2016). https://doi.org/10.1007/s00521-016-2686-9
19. Kannappan, P.L.: On Shannon's entropy, directed divergence and inaccuracy. Probab. Theory Rel. Fields **22**, 95–100 (1972). https://doi.org/10.1016/S0019-9958(73)90246-5
20. Paul, S., Bandyopadhyay, B.: A novel approach for image compression based on multi-level image thresholding using Shannon entropy and differential evolution, In: IEEE Students' Technology Symposium (TechSym), pp. 56–61 (2014). https://doi.org/10.1109/techsym.2014.6807914
21. Rajinikanth, V., Satapathy, S.C., Fernandes, S.L., Nachiappan, S.: Entropy based segmentation of tumor from brain mr images—A study with teaching learning based optimization. Pattern Recogn. Lett. **94**, 87–94 (2017)
22. Shanthakumar, P., Kumar, P.G.: Computer aided brain tumor detection system using watershed segmentation techniques. Int. J. Imaging Syst. Technol. **25**(4), 297–301 (2015). https://doi.org/10.1002/ima.22147
23. Rajinikanth, V., Raja, N.S.M., Kamalanand, K.: Firefly algorithm assisted segmentation of tumor from brain MRI using Tsallis function and Markov random field. J. Control Eng. Appl. Inform. **19**(3), 97–106 (2017)
24. Lu, H., Kot, A.C., Shi, Y.Q.: Distance-reciprocal distortion measure for binary document images. IEEE Signal Process. Lett. **11**(2), 228–231 (2004)
25. Moghaddam, R.F., Cheriet, M.: A multi-scale framework for adaptive binarization of degraded document images. Pattern Recogn. **43**(6), 2186–2198 (2010)
26. Radiopedia: Sub-acute middle cerebral artery infarct database—Case courtesy of Dr. David Cuete, Radiopaedia.org, rID: 35732
27. ISLES: (2015). www.isles-challenge.org
28. ITK-SNAP: http://www.itksnap.org/pmwiki/pmwiki.php

Scheme for Unstructured Knowledge Representation in Medical Expert System for Low Back Pain Management

Debarpita Santra, Sounak Sadhukhan, S. K. Basu, Sayantika Das, Shreya Sinha and Subrata Goswami

Abstract A fundamental requirement to achieve effectiveness in a medical expert system is the proper representation of knowledge about a patient. Knowledge about a patient is stored in the expert system in terms of clinical records which should contain information about multiple visits of the patient. During each visit, several conversations are made between the patient and the consulting physician. These conversations, being unstructured in nature, cannot be stored in the computer using available structured knowledge representation schemes. So, we propose a recursive frame-based structure for representing clinical records. The frames related to a patient collectively form a frame system, where one frame may point to other frames. The proposed representation scheme is complete, consistent, and free from redundancy.

D. Santra · S. Sadhukhan (✉) · S. K. Basu
Department of Computer Science, Banaras Hindu University, Varanasi 221005, India
e-mail: sounaks.cse@gmail.com

D. Santra
e-mail: debarpita.cs@gmail.com

S. K. Basu
e-mail: swapankb@gmail.com

S. Das · S. Sinha · S. Goswami
ESI Institute of Pain Management, Kolkata 700009, India
e-mail: 93sayantika@gmail.com

S. Sinha
e-mail: shreyasinha256@gmail.com

S. Goswami
e-mail: drsgoswami@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_4

1 Introduction

Medical diagnosis is made based on clinical tests and clinical examinations by the physician. There has been fast and tremendous growth of knowledge and advances in medical and other associated techniques. It is quite difficult and almost impossible for a treating physician to be at home with all these advances and pass on the benefits to the patients. A medical expert system is defined as a software using artificial intelligence techniques and up-to-date knowledge of a specific medical domain. Advantages of medical expert system can obviously be appreciated at the initial screening stages of diseases like low back pain (LBP).

Compared to any other diseases, LBP is responsible for causing more disability worldwide [1]. Though several methods are used in clinical practice to investigate and diagnose LBP, unfortunately, the exact cause of producing the LBP is not well identified sometimes if the patient is not subjected to elaborate clinical investigations. To cope with this issue, we are planning to design a reliable and efficient medical expert system to assess and manage LBP.

Like any expert system, the medical expert system for LBP should consist of four major modules: user interface, working memory, knowledge base, and inference engine. The patient information that are gathered through the user interface, should be represented and stored in the expert system in such a way that they can be interpreted and retrieved efficiently. This kind of information is highly unstructured in nature. So, capturing the semantics as well as the complex relationships among the data items is a challenging task. In this paper, we have proposed a frame-based representation scheme for the unstructured LBP patient information.

The rest of the paper is organized as: Sect. 2 gives an overview of related works. In Sect. 3, we discuss our proposed representation scheme and illustrate it using a simple example from the domain of LBP. Finally, Sect. 4 concludes the paper.

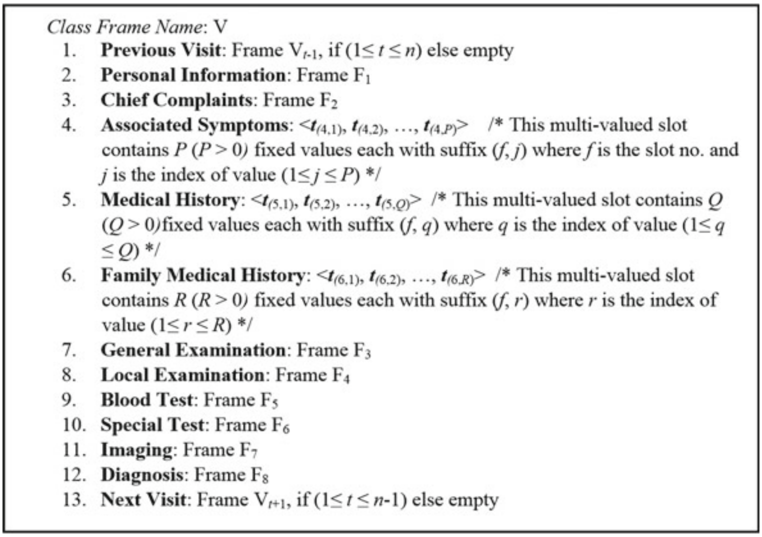
2 Related Study

Efforts have been made since 1970s to build up medical expert systems. MYCIN [2] which is the pioneering medical expert system for blood infections, aids physicians to recommend antibacterial medicines for the disease. Other renowned medical expert systems are CASNET [3] for glaucoma, INTERNIST [4] applicable for the domain of internal medicines, ONCOCIN [5] for cancer, etc. Some expert systems [6–8] are concerned with LBP. In case of MYCIN, a context tree is formed to hold the previous medical histories of patients and each instance of the tree basically represents individual clinical event [9]. This representation is not effective when a clinical record is visualized as a collection of subsequent events or visits. CASNET is a causal model for better understanding of disease processes. The findings about a patient are represented as a collection of individual nodes in the plane of observation, but the complex relationships among patient findings are not well captured. INTERNIST proposes a hierarchical organization for different disease categories that are associated with different organ systems in a human body. Each of the disease categories is further subdivided until leaf nodes are reached. A leaf node represents an individual disease. ONCOCIN proposes a temporal network to obtain the key information about the past chemotherapy cycle or previous visits of a patient. But this structure is also not efficient to capture the interrelatedness of patient data that appear in the domain of LBP [10–13] did not propose any unstructured knowledge representation scheme for capturing complex relationships among data items in individual patient records.

3 Proposed Scheme

A patient record should comprise all the information that are collected during multiple visits of the patient. During these visits, lots of conversations take place between the patient and the consulting physician. These conversations are fully unstructured in nature and cannot be represented in computer using the available representation schemes for structured data. Here, we have proposed a suitable representation scheme for patient record to capture and analyze important relationships that exist among data items. We represent each visit as an individual frame that can capture different types of unstructured data with the help of attributes (slots) and values (fillers) associated with the visit. A slot may hold multiple values or no value at all. A frame allows recursive embedding, i.e., it can have other frames as slots. We visualize a patient record as a sequential collection of information related to visits. So, a patient record can be represented as a frame system which is a collection of connected frames. This representation scheme would facilitate achieving computational effectiveness of the medical expert system for LBP.

A frame associated with the tth visit ($1 \le t \le n$) is represented as V_t. So, a complete patient record R is a frame system with a frame V_t pointing to another frame V_{t+1}. We now use the concept of a class frame which can have many instances. We assume that there exist subclass frames of a class frame, where each subclass frame contains less or equal number of attributes of the class frame. A subclass frame may also have many instances. We define a class frame V as a collection of L slots ($L > 0$), where L_1 ($L_1 > 0$) slots contain only fixed values, L_2 (=2) slots contain two different frame instances of V, one for previous visit (if any) and another for the next visit (if any). The remaining slots L_3 ($= L - (L_1 + L_2)$) contain frames different from frames of type V. An individual slot from L_3 contains a particular frame F_i ($1 \le i \le L_3$), which is defined as a collection of M slots ($M > 0$), where M_1 ($M_1 > 0$) slots contain only fixed values and M_2 ($= M - M_1$) slots contain another F_j ($j > L_3$) frames. The values of M, $M_{1,}$ and M_2 as well as the attributes will be different for each F_i and F_j frames. A class frame V may have many subclasses. A subclass frame V^s of V contains slots equal or less than that of V, but will always contain all the L_2 slots. The attributes that will be present in subclass frame V^s at tth visit would be decided dynamically based on the conversation on that visit. So, we say that V_t is an instance of V^s. We provide the generalized definition of the class frame V in Fig. 1.



Class Frame Name: V

1. **Previous Visit**: Frame V_{t-1}, if ($1 \le t \le n$) else empty
2. **Personal Information**: Frame F_1
3. **Chief Complaints**: Frame F_2
4. **Associated Symptoms**: <$t_{(4,1)}$, $t_{(4,2)}$, …, $t_{(4,P)}$> /* This multi-valued slot contains P ($P > 0$) fixed values each with suffix (f, j) where f is the slot no. and j is the index of value ($1 \le j \le P$) */
5. **Medical History**: <$t_{(5,1)}$, $t_{(5,2)}$, …, $t_{(5,Q)}$> /* This multi-valued slot contains Q ($Q > 0$)fixed values each with suffix (f, q) where q is the index of value ($1 \le q \le Q$) */
6. **Family Medical History**: <$t_{(6,1)}$, $t_{(6,2)}$, …, $t_{(6,R)}$> /* This multi-valued slot contains R ($R > 0$) fixed values each with suffix (f, r) where r is the index of value ($1 \le r \le R$) */
7. **General Examination**: Frame F_3
8. **Local Examination**: Frame F_4
9. **Blood Test**: Frame F_5
10. **Special Test**: Frame F_6
11. **Imaging**: Frame F_7
12. **Diagnosis**: Frame F_8
13. **Next Visit**: Frame V_{t+1}, if ($1 \le t \le n$-1) else empty

Fig. 1 Generalized structure of class frame V for the domain of LBP

For our current clinical consideration about the structure of V, $L = 13$, where $L_1 = 3$ and $L_3 = 8$. The attributes of frames F_1, F_2,…, F_8 (Fig. 1) are different from the medical perspective. For example, while F_1 captures personal information like name, age, sex, occupation, etc., of an LBP patient as fixed values, the frame F_2 captures the chief (major) complaints of the patient like the duration and type of pain, radiation (if any), aggravating factors, relieving factors, diurnal variation, etc. Each of the complaints is captured using different slots of the frame. While some of the slots may contain fixed values (if not empty), other slots indicating radiating pain, aggravating factors, relieving factors, and diurnal variation point to other frames F_9 through F_{12}, respectively. These frames may also have recursive embedding. We illustrate our proposed scheme using an example as shown in Fig. 2.

For the sake of illustration of our proposed scheme using Fig. 2, we have considered a patient named XX who is a jute mill worker. The patient has visited the doctor for the first time on December 26, 2016 with complaints of LBP. The patient is a female of age 56 years. She feels continuous pain below L5 of the spine and in buttocks since last 2 months. Most of the time she becomes depressed with her work and her discomfort level is 6 on the Visual Analog Scale (VAS). So, the severity of her pain can be interpreted as “high”. The pain starts slowly and is aggravated when she bends or lifts some moderately heavy object. Also, her pain is aggravated when she is sitting or standing for long time. Her pain is aggravated in morning and night also. Her pain radiates to the right lower leg at every night. Her sleep is disturbed and bowel/bladder habit is normal. Associated with this, she also has symptoms of headache and nausea. She has allergy and suffers from diabetes. Her mother had died of cancer recently and her father is diabetic.

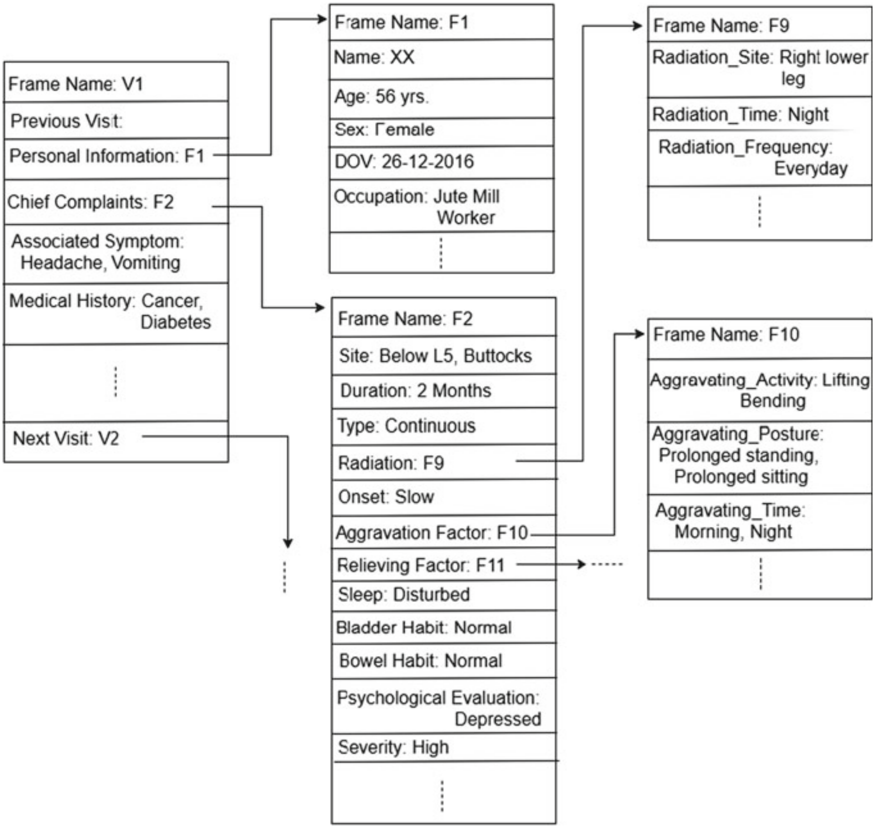


Fig. 2 Representation of our example using the proposed scheme

3.1 *Formal Representation of a Patient Record*

We have formally represented a record of a patient P using six tuples <N_p, T_p, R_p, P_p, C_p, D_p>, where

- N_p is a set of variables and $N_p = \boldsymbol{V} \cup \boldsymbol{F} \cup \boldsymbol{A}$, where $\boldsymbol{V} = \{V_t \mid V_t$ is a frame variable corresponding to the frame V_t $(1 \le t \le n)\}$, $\boldsymbol{F} = \{F_{(t,m)} \mid F_{(t,m)}$ is a frame variable corresponding to the frame F_m at *t*th visit and total no. of F_m variables is greater than or equal to $L_3\}$, $\boldsymbol{A}$ is the set of temporary variables required for ease of internal computations.
- T_p is a set of fixed values in the frame system, where $T_p = T_{\boldsymbol{V}} \cup T_{\boldsymbol{F}}$, where $T_{\boldsymbol{V}}=\{\boldsymbol{t}_{(a,b),t} \mid \boldsymbol{t}_{(a,b),t}$ is the *b*th $(b>0)$ fixed value of the slot a $(1 \le a \le L_1)$ for frame $V_t\}$, $\boldsymbol{T_F} = \{\boldsymbol{t}_{k,(t,i)} \mid \boldsymbol{t}_{k,(t,i)}$ $(1 \le k \le K)$, K being the total number of fixed values that frame F_i can contain at *t*th visit$\}$
- R_p is called the starting point where $R_p \in \boldsymbol{V}$

- P_{p} is a set of replacement rules for retrieving the record of a patient in terms of fixed values finally. A replacement rule is of the form $\alpha \rightarrow \beta$, where $\alpha \in N_{\mathrm{p}}$ and β represents strings on $(N_{\mathrm{p}} \cup T_{\mathrm{p}})^*$. The set of replacement rules is given below:

  (i) $R_p \rightarrow V_t$

  (ii) $V_t \rightarrow V_{t-1}\, A_1 \mid A_1$/* $A_1 \in \boldsymbol{A}$ */

  (iii) $A_1 \rightarrow (\wedge\, F_{(t,i1)})\, (\wedge\, \boldsymbol{t}_{(a,b),t})\, (\wedge\, F_{(t,i2)}) \mid (\wedge\, \boldsymbol{t}_{(a,b),t}) \mid F_{(t,1)}$
  /* (for all a, b) and with $1 \le i1 \le 2$ and $i2 > i1$, $(i1 + i2) = L_3$, and $(\wedge\, F_{(t,i)})$, $(\wedge\, \boldsymbol{t}_{(a,b),t})$ are concatenation of subsequent frames and concatenation of subsequent fixed values, respectively */

  (iv) $F_{(t,i)} \rightarrow TA_2T \mid TA_2 \mid A_2\, T \mid T \mid \lambda$
  /* λ means Null value and $i1$, $i2$ from rule (iii) and $i3$ from rule (v) will be replaced by i. Here, T, $A_2 \in \boldsymbol{A}$ */

  (v) $A_2 \rightarrow F_{(t,i3)}\, A_2 \mid \lambda$
  /* $i3$ represents the subsequent indices of pointed frames of $F_{(t,i)}$ and is different from $i1$ and $i2$*/

  (vi) $T \rightarrow \boldsymbol{t}_{k,(t,i)}\mathrm{T} \mid \boldsymbol{t}_{k,(t,i)}$

- C_p is the set of constraints imposed on invoking of the replacement rules. Constraints define on which conditions or contexts a replacement rule should be invoked. Constraints also define when to terminate in case of recursive calls. Also, constraints impose restrictions on sequence of invocation of replacement rules to derive a final string of fixed values.
- D_p is the set of dependencies among a set of frame variables and a set of fixed values. An element (either variable or fixed value) may depend on another element or set of elements to aid in deducing the final string that represents a partial or complete patient record. For example, the value of the frame slot “Diagnosis” in V_t would depend on the values of all previous frames and fixed values in that frame.

3.2 Reasoning Mechanism

Our proposed representation scheme does not allow random or direct access of a fixed value of a frame. That is, we have to follow a systematic pathway to get the information connected with the tth visit. After all the internal processing of frame V_1 is complete, we would proceed to the next visit (frame V_2) and in this way, we will step forward sequentially until we reach frame V_t. We can construct a derivation tree using the replacement rules we have provided in Sect. 3.1. The root of the tree is R_{p}. Every node v of the tree is either a variable or a fixed value or λ. If we have a replacement rule such as $X \rightarrow X_1\ X_2\ X_3$, where all three at R.H.S. of the rule are variables, we say that X will be placed at a parent node v in the tree and X_1, X_2, X_3 will be placed at the child nodes v_1, v_2, v_3 of v, respectively. Now, suppose we have another rule $X_1 \rightarrow Y_1\ Y_2$, where Y_1 and Y_2 are fixed values. So, node v_1 will have two children $v_{1.1}$ and $v_{1.2}$ which are leaves. We also consider another replacement rule $X_2 \rightarrow Y_3\ Y_4\ Y_5$, where in R.H.S., first two are variables, and the last one is a fixed value and so on. The access/reasoning sequence will be X- X_1- Y_1- Y_2- X_2- Y_3- Y_4- Y_5- X_3… So, always the leftmost variable present on the R.H.S. of a replacement rule $\boldsymbol{H}$ will be explored first until it reaches the leaf nodes, then the next variable of $\boldsymbol{H}$ will be explored. This process will continue until we visit all the leaves of the tree. The patient record that is the yield of the derivation tree is the concatenation of the leaf nodes that appear in the order from left to right with no repetition. If we want to retrieve the information (INFO) corresponding to a slot $\text{Slot}_{k,(t,i)}$ in an individual frame $F_{(t,i)}$, we simply use syntax <$\text{Slot}_{k,(t,i)}$.$\text{INFO}_{k,(t,i)}$>. Let us illustrate the reasoning mechanism with our example.

R_p => V_1 // using rule (i): $R_p \rightarrow V_t$
=> A_1 // using rule (ii): $V_t \rightarrow A_1$
=> $F_{(1,1)}\ F_{(1,2)}$<Associated Symptom.(Headache, Vomiting)>**…**
// using rule (iii): $A1 \rightarrow (\wedge F_{(t,i1)})(\wedge \boldsymbol{t}_{(a,b),t})(\wedge F_{(t,i2)})$
=> $T\ F_{(1,2)}$ < Associated Symptom.(Headache, Vomiting)>**…**
// using rule (iv): $F_{(t,i)} \rightarrow T$
=> <Name.XX><Age.56 yrs><Sex.Female>**…**<DOV.26-12-2016>**…**
<Occupation.Jute Mill Worker>**…** $F_{(1,2)}$<Associated Symptom.(Headache, Vomiting)>**…**
// using rule (vi): $T \rightarrow \boldsymbol{t}_{k,(t,i)}T \mid \boldsymbol{t}_{k,(t,i)}$
=> <Name.XX><Age.56 yrs><Sex.Female>**…**<DOV.26-12-2016>**…**
<Occupation.Jute Mill Worker>**…** TA_2T <Associated Symptom.(Headache, Vomiting)>**…**
// using rule (iv): $F_{(t,i)} \rightarrow TA_2T$
=> <Name.XX><Age.56 yrs><Sex.Female>**…**<DOV.26-12-2016> **…**
<Occupation.Jute Mill Worker>**…**<Site.(Below L5, Buttock)><Duration.2 months><Type.Continuous>$F_{(1,9)}A_2T$ <Associated Symptom.(Headache, Vomiting)>**…**
// using rule (vi): $T \rightarrow \boldsymbol{t}_{k,(t,i)}T$ multiple times
=> <Name.XX><Age.56 yrs><Sex.Female>**…**<DOV.26-12-2016>**…**
<Occupation.Jute Mill Worker>**…**<Site.(Below L5, Buttock)><Duration.2 months><Type.Continuous> <Radiation_Site.Right lower leg>
<Radiation_Time.Night><Radiation_Frequency.Everyday>A_2T
<Associated Symptom.(Headache, Vomiting)>**…**
// using rule (iv): $F_{t,i} \rightarrow T$ and rule (vi): $T \rightarrow \boldsymbol{t}_{k,(t,i)}\ \boldsymbol{T}$ multiple times
=> **…….**
=> <Name.XX><Age.56 yrs><Sex.Female>**…**<DOV.26-12-2016>**…**
<Occupation.Jute Mill Worker>**…** <Site.(Below L5, Buttock)><Duration.2 months><Type.Continuous><Radiation_Site.Right lower leg><Radiation_Time.Night><Radiation_Frequency.Everyday>
<Aggravating_Activity.(Lifting, Bending)>
<Aggravating_Posture.(Prolonged Sitting, Prolonged Standing)>
<Aggravating_Time.(Morning, Night)>**…**<Associated Symptom.(Headache, Vomiting)>**…**
// using rule (vi): $T \rightarrow \boldsymbol{t}_{k,(t,i)}T \mid \boldsymbol{t}_{k,(t,i)}$ multiple times

The final string is the concatenation of all the fixed values that represents the overall information of the patient during her visit.

3.3 Discussion of Properties

Our proposed scheme has a number of characteristics as follows. It is complete from two aspects. Firstly, we are able to capture the entire record of a patient visitwise, and secondly, we can reach at fixed values starting from R_{p} without entering into infinite loop. The scheme is consistent, as we have defined range of all the variables and also defined what kind of value a variable may be assigned. We also keep provision of storing null values. Every variable is different at every visit and if a variable accepts two different fixed values at two different visits, it will not be a consistency violation, as diagnosis is dependent on the latest value.

Moreover, the structure works in an integrated manner, which is ensured by recursive embedding of frames in an individual frame. Also, there are dependencies among a set of variables and a set of fixed values during a particular visit. This also ensures the integrity of our scheme. The proposed structure does not support redundant information, as we have partitioned our total structure into many frames, where each frame contains a particular type of information. Also, when a particular frame points to other frames, we distinguish the other frames using different indices. So, a frame does not point to itself, which ensures that our design is free from redundancy. For easy access to the partial or complete patient record, we differentiate multiple visits using variable t and use simple syntax <V_t.Record> to grab the information till the tth visit.

4 Conclusion

In this paper, we have indicated how frame-based knowledge representation scheme provides the foundation for describing a patient record, which contains unstructured data items also. To make the representation clinically intuitive and computationally efficient, we have used the concept of recursive embedding in a single frame. We have also explained how to reason using this representation. In the context of diagnosing LBP, the design has the properties like completeness, consistency, integrity, and non-redundancy. The design facilitates ease of access of the stored data items. We plan to use this scheme for implementation of medical expert system for LBP management in the near future.

References

1. WebMD: https://www.webmd.com
2. Shortliffe, E.H.: Computer-Based Medical Consultation: MYCIN (1976)
3. Weiss, S.M., Kulikowski, C.A., Safir, A.: A model-based consultation system for the long term management of glaucoma. In: 5th International Joint Conference on Artificial Intelligence, pp. 826–832 (1977)
4. Pople Jr., H.E.: The Formation of composite hypotheses in diagnostic problem solving: an exercise in synthetic reasoning. In: 5th International Joint Conference Artificial Intelligence, pp. 1030–1037 (1977)
5. Kahn, M.G., Ferguson, J.C., Shortliffe, E.H., Fagan, L.M.: Representation and use of temporal information in ONCOCIN. In: Annual Symposium on Computer Application in Medical Care, pp. 172–176 (1985)
6. Toth-Tascau, M., Stoia, D.I., Andrei, D.: Integrated methodology for a future expert system used in low back pain management. In: 7th IEEE International Symposium on Applied Computational Intelligence and Informatics, pp. 315–320 (2012)
7. Abu Naser, S., AlDahdooh, R.: Lower back pain expert system diagnosis and treatment. J. Multi. Eng. Sci. Stud. **2**, 441–446 (2016)
8. Landi, A., Davis, R., Hendler, N., Tailor, A.A.: Diagnoses from an on-line expert system for chronic pain confirmed by intra- operative findings. J. Anesth. Pain Med. **1**, 1–7 (2016)
9. Shortliffe, E.: Computer-Based Medical Consultations: MYCIN (2012)
10. Kong, G., Xu, D., Yang, J.: Clinical decision support systems: a review on knowledge representation and inference under uncertainties. Int. J. Comput. Intell. Syst. **1**, 15–167 (2008)
11. Sundari, A.M., Balasundaram, R.: Computerisation in obstetrics and gynaecology—An expert approach. In: IEEE Engineering in Medicine and Biology Society and 14th Conference of the Biomedical Engineering Society of India, pp. 8–9 (1995)
12. Klar, R., Zaiss, A.: Medical expert systems: design and applications in pulmonary medicine. Lung **168**(Suppl), 1201–1209 (1990)
13. Barnett, G.O., Hoffer, E.P., Packer, M.S., Famiglietti, K.T., Kim, R.J., Cimino, C., Feldman, M.J., Oliver, D.E., Kahn, J.A., Jenders, R.A., Gnassi, J.A.: DXplain—Demonstration and discussion of a diagnostic decision support system. In: Annual Symposium on Computer Application in Medical Care, pp. 822 (1993)

An Unvarying Orthogonal Search with Small Triangle Pattern for Video Coding

S. Immanuel Alex Pandian and J. Anitha

Abstract In video coding systems, motion estimation consumes more time due to its complexity. A new search pattern has been developed in this paper to decrease the complexity in computing the motion estimation in block matching algorithm. This algorithm basically employs an orthogonal search pattern with uniform step size. A fast inner search is applied by choosing a small triangle pattern adjacent to the region with the minimum group sum distortion that further reduces the search points. Experimental results demonstrate the significance of the proposed approach in the reduction of the search point computation by 28.59–84.69% over existing fast motion estimation approaches thus maintaining the image quality.

1 Introduction

Motion estimation takes an inevitable part in many applications of video compression. Approximately 60–80% of the total computation is taken by the motion estimation, so it has a substantial result on the video coding performance. Block-based motion estimation will estimate the motion of blocks from frame to frame. It divides the frames into Macro Blocks (MBs). In the current frame, each MB is matched against the MB in the reference frame, surrounded by a search window depicted in Fig. 1. The value of search parameter "d" is proportional to the motion in the video sequences.

Various Block Matching Algorithms (BMAs) have been established to achieve the computational complexity reduction. Motion estimation through Full Search (FS)

S. Immanuel Alex Pandian
Department of ECE, Karunya Institute of Technology and Sciences, Coimbatore,
Tamil Nadu, India
e-mail: immans@karunya.edu

J. Anitha (✉)
Department of CST, Karunya Institute of Technology and Sciences, Coimbatore,
Tamil Nadu, India
e-mail: anitha_j@karunya.edu

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_5

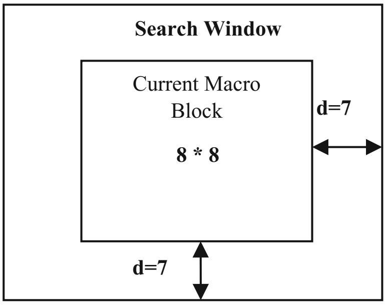


Fig. 1 A macroblock with search parameter "d" of ±7 pixels and size of 8 × 8 pixels

is simple and it finds the finest match and provides good accuracy [1]. But due to the computation of all available macro blocks inside the search window, it takes high complexity. Several fast algorithms for block matching such as Three Step Search (TSS) [2], New Three Step Search (NTSS) [3], Four Step Search (FSS) [4], Diamond Search (DS) [5], Orthogonal Search (OS) [6], Hexagonal Search (HS) [7], Star Diamond (SD) [8], and many other algorithms [9–13] have been proposed to reduce the search point computation. Further, a fast inner search is highly desirable to decrease the search points. The expected inner search performs an extensive full search in the small area. Several fast algorithms [14, 15] have been developed that focus on reducing the search points by finding efficient patterns for inner search.

This paper proposes an unvarying orthogonal search with group-sum distortion to select small triangle pattern for the inner search that can effectively reduce the search point computations. In conventional orthogonal search [6], first the search is performed in vertical stage followed by the horizontal stage. Each step, the step size is halved until the step size equals one. But this pattern is inefficient for small motions because it might get stuck into local minimum. In this proposed algorithm, the same search pattern as orthogonal is taken with initial step size set to two and later with no variation of step size in each iteration. Then the inner search pattern is selected near to the group, based on group-sum distortion of the four points surrounded the center point with minimum distortion.

The content of this paper is structured in the following order. Section 2 introduces the proposed unvarying orthogonal search with small triangle search pattern algorithm. Section 3 describes simulation results and comparative analysis of the proposed method with existing block matching algorithms. Finally, Sect. 4 concludes this paper.

2 Proposed Algorithm—UOSSTP

The above discussed fast block matching algorithms utilized different search patterns that provide less search points computation for finding the finest motion vector. This paper proposes an Unvarying Orthogonal Search with a Small Triangle Pattern

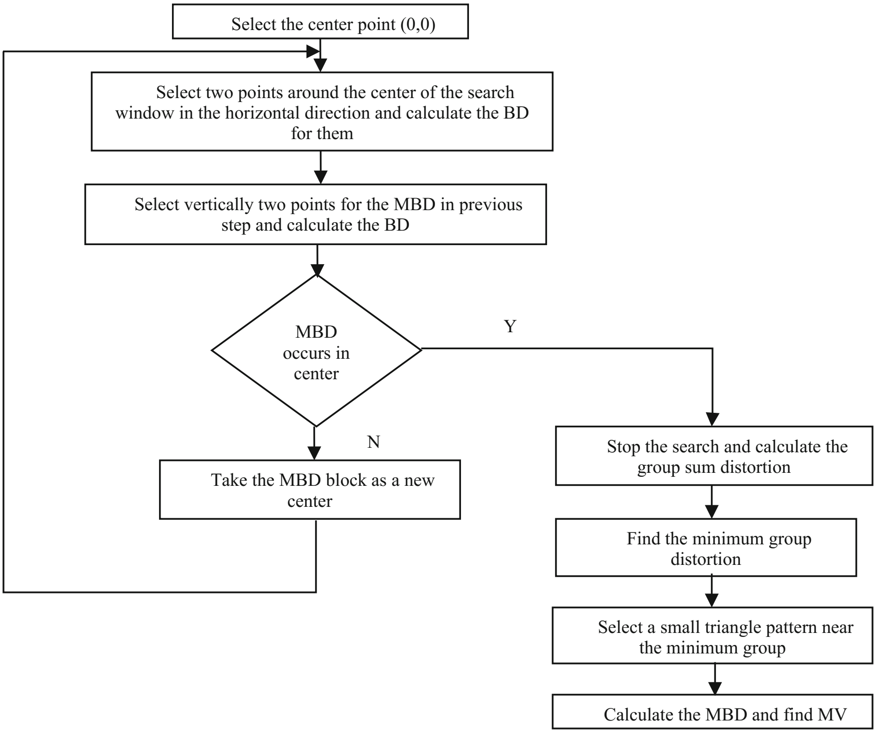

Select the center point (0,0)
Select two points around the center of the search window in the horizontal direction and calculate the BD for them
Select vertically two points for the MBD in previous step and calculate the BD
MBD occurs in center
Y
N
Take the MBD block as a new center
Stop the search and calculate the group sum distortion
Find the minimum group distortion
Select a small triangle pattern near the minimum group
Calculate the MBD and find MV

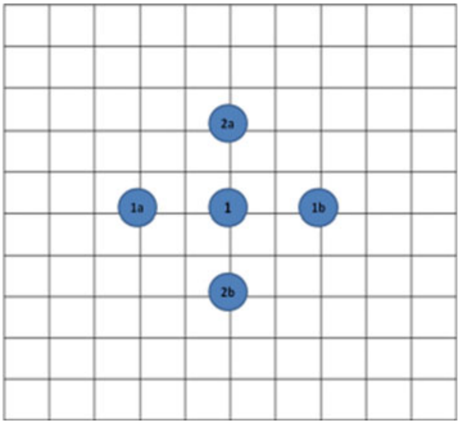
Fig. 2 Framework of the UOSSTP motion estimation algorithm

(UOSSTP) to enhance the process of estimating the block motion. The framework of the UOSSTP motion estimation algorithm is depicted in Fig. 2.

2.1 Unvarying Orthogonal Search Pattern

A statistical result suggests that 40–70% best optimal point is positioned around the center point. Many BMAs are inspired by the center-biased motion vector nature [5–7]. So the proper design and selection of search strategy and search pattern can ensure the speed up in motion estimation. This unvarying orthogonal search pattern starts the search around the center point within the search window follows the procedure of vertical direction followed by the horizontal direction. This search contains three search points with the center enclosed by two horizontal search points (1a, 1b) from the center point with the distance of 2. The point with minimum block distortion is selected as center for next step. Two points in vertical direction (2a, 2b) are taken around the new center. Then in the vertical direction, the new center is

Fig. 3 Unvarying orthogonal search pattern



calculated. This process continues until the minimum block distortion to be obtained is the center point itself.

Figure 3 depicts the initial step of the unvarying orthogonal search pattern. To improve the efficiency further, a new inner search pattern is introduced near the center with minimum group distortion that handles small motions.

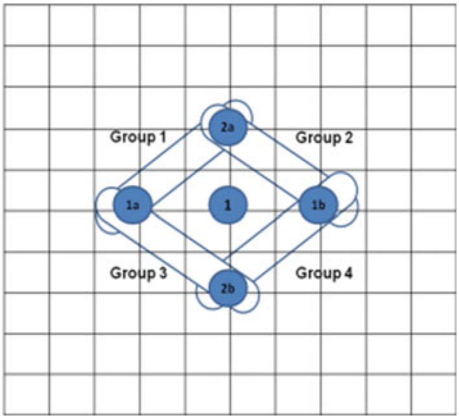
2.2 Small Triangle Search Pattern

After the center point with minimum block distortion (MBD) is identified in the unvarying orthogonal search pattern, a fast inner search is performed to attain further search point reduction in overall. The distortion information of the four search points around the center with minimum distortion in the coarse level is used for the inner search.

Figure 4 depicts the grouping of horizontal and vertical search points around the center. For every individual group, a group distortion is evaluated by adding the block distortion measure of points present in that group. The region nearby the group with less group block distortion is marked as a minimum distortion found region. Four different groups are formed by combining four search points around the center of unvarying orthogonal search pattern. Group 1 is formed by combining points 1a and 2a. Group 2 is formed by combining points 2a and 1b. Group 3 is formed by combining points 1a and 2b. Finally, Group 4 is formed by combining points 1b and 2b.

The focused inner search uses a small triangle pattern of three search points in the region nearer to the minimum distortion group. So, this extra point selection is mainly based on the horizontal and the vertical search points selected from the uniform orthogonal search pattern. The overhead of performing these four addi-

Fig. 4 Group sum of unvarying orthogonal search pattern



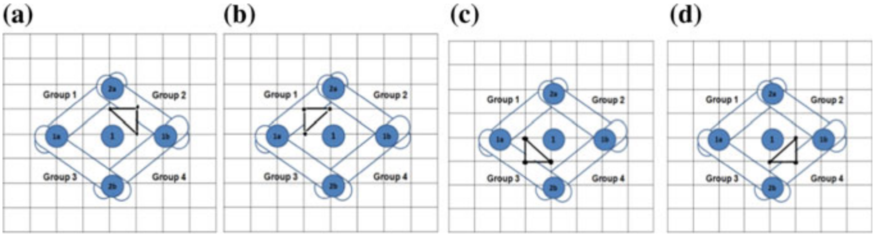


Fig. 5 Inner search patterns for UOSSTP. **a**–**d** Small triangle patterns based on group-sum distortions

tions to calculate the group sum distortion is negligible. The proper usage of nearby information would produce an accurate and yet faster inner search.

In Fig. 5, if the horizontal search point (1a) and the vertical search point (2a) form Group 1, the algorithm utilizes the small triangle seen in Fig. 5a, if Group 2, it uses Fig. 5b, if Group 3, it uses Fig. 5c, if Group 4, it uses Fig. 5d. In case of Enhanced Hexagonal Search (EHS), the amount of points to be searched in the inner search is not same for all group selections. Depending on the location of the group, the inner search can take either two or three search points. When compared to the small diamond pattern in DS and small hexagon pattern in HS, the proposed method uses a small triangle pattern for inner search. This pattern uses only three search points and can attain more improved results.

2.3 *Search Point Analysis*

Figure 6 depicts the search point analysis for the MBD with center point is not (0, 0). The per block search points derived from search point analysis is given in Eq. 1 as

Fig. 6 Search path example with MBD is not (0, 0)

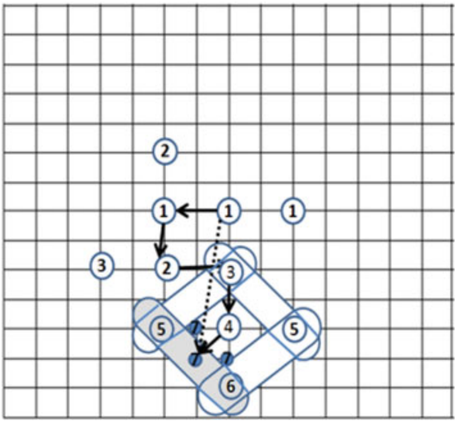


Table 1 The video sequences used for the implementation

Test sequences	Video format	Frame size	Motion type
Container	Quarter CIF	176 × 144	Small
Akiyo	Quarter CIF	176 × 144	Small
Carphone	Quarter CIF	176 × 144	Medium
Foreman	Quarter CIF	176 × 144	Medium
Miss-America	Quarter CIF	176 × 144	Medium
Salesman	Quarter CIF	176 × 144	High
Stefan	CIF	352 × 288	High
Tennis	SIF	352 × 240	High

$$N_{\mathrm{UOSTS}} = 5 + M * n + 3 \tag{1}$$

where M is either 4 or 3 depending on the points to be checked in horizontal and vertical direction, and n represents the number of execution.

Equation (1) clearly shows that the coarse level search of UOSSTP starts with only 5 points. Also, the fine level search requires 3 points where the intermediate search requires 3–4 points for UOSSTP based on the number of iterations (n).

3 Results and Discussions

The standard test video sequences in which the experiments are carried out are listed in Table 1. The results obtained through the experiments are compared against the available state-of-the-art methods include FS, OS, DS, and HS. The block distortion criterion named Mean Absolute Difference (MAD) is used to estimate the best similar block in this work.

Table 2 The computed per block average search points for various BMAs and various sequence of video

	Container	Akiyo	Carphone	Foreman	Miss-America	Stefan	Tennis	Salesman
FS	191.1313	191.1313	191.1313	191.1313	191.1313	191.1313	191.1313	191.1313
DS	12.6415	11.8643	13.8162	13.6768	13.8255	16.8336	16.4319	12.1980
OS	12.0102	10.9981	12.0545	11.9992	12.0782	12.5434	12.4964	11.9996
HS	10.3602	9.9554	10.9825	10.9239	11.1540	12.0443	12.2705	10.1400
UOSSTP (proposed)	8.0261	7.7324	8.3979	8.2017	8.4753	9.1146	9.5418	7.8227

These test video sequences contain different types of motion and different types of format include Source Input Format (SIF), Common Intermediate Format (CIF), and Quarter Common Intermediate Format (QCIF). The experiment uses the macroblock of size 8 × 8 with the first 100 frames of each sequence. The displacement range in vertical and horizontal direction for a block is considered as $d = \pm 7$. The experiments are implemented using the MATLAB software.

The computed per block average search points for various BMAs are tabulated in Table 2. The outcomes evident that the developed UOSSTP algorithm requires fewer points to be searched which in turn reduces the computations as compared to other available BMAs. The obtained per block average search points from the results are, UOSSTP < HS < OS < DS < FS.

Table 3 demonstrates the performance on PSNR for different test sequences with different methods. PSNR is relatively low for Stefan and Tennis which are large motion sequences. Akiyo, Container, and Miss-America are slow-motion sequences where the PSNR value is relatively high. The result shows that there is an improvement in PSNR compared to other fast ME approaches except FS, which indicates that the proposed algorithm does not compromise the accuracy.

The Speed Improvement Rate (SIR) on average is summarized in Table 4 for various competing algorithms. The SIR in terms of % is represented by Eq. (2) as

$$\mathrm{SIR} = (N_2 - N_1)/N_1 \tag{2}$$

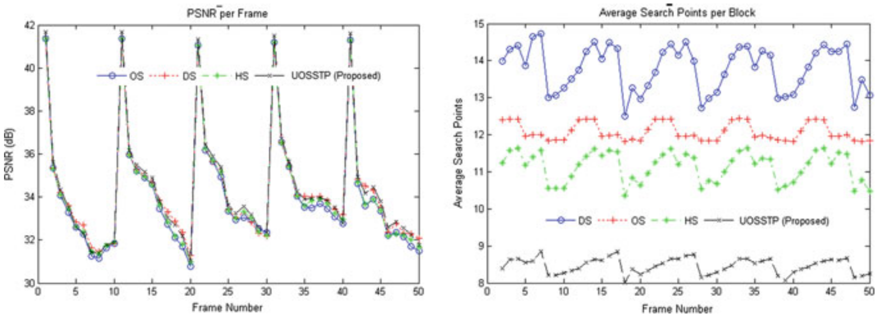
where N_1 denotes the number of total points to be searched in Method1, while N_2 denotes the number of total points to be searched in Method2.

For *Container* sequence with small motions, the SIR of UOSSTP over HS is 29.08%, UOSSTP over DS is 57.5% and UOSSTP over OS is 49.64%. For large motion, UOSSTP algorithm has attained speed improvement of 85% over other algorithms.

Figure 7 plots the comparative analysis of frame-by-frame evaluation on *Miss-America* sequence of the first 50 frames to represent PSNR and per block average search points respectively related to the various BMAs. This figure shows an improved reduction in search points with acceptable image quality.

Table 3 The mean PSNR computed for different BMAs and various series of videos

	FS	OS	DS	HS	UOSSTP (proposed)	Comparison		
						ΔOS	ΔDS	ΔHS
Container	33.4465	33.4360	33.4365	33.4339	33.4405	0.0045	0.0040	0.0066
Akiyo	36.3943	36.1379	36.3652	36.2326	36.3816	0.2437	0.0164	0.1490
Carphone	30.9647	30.2670	30.6808	30.2806	30.7588	0.4918	0.0780	0.4782
Foreman	30.3759	28.8765	29.7462	29.1106	29.9193	1.0428	0.1731	0.8087
Miss-America	34.4434	34.1314	34.3582	34.2094	34.3980	0.2666	0.0398	0.1886
Stefan	24.8443	23.5215	23.6598	23.4255	23.9698	0.4483	0.3100	0.5443
Tennis	28.1500	26.7894	27.5564	27.3223	27.9079	1.1185	0.3515	0.5856
Salesman	32.8520	32.5340	32.6990	32.5805	32.7488	0.2148	0.0498	0.1683



(a) Average PSNR per frame

(b) Average number of search points

Fig. 7 Comparative analysis of DS, OS, HS and the proposed UOSSTP to represent PSNR per frame and the per block average search points, respectively for *Miss-America* video sequence

Based on the results obtained through the experiment, it is clearly seen, the proposed search yields fewer points to be searched than OS, DS, and HS algorithm while maintaining good performance in image quality. As an outcome, the algorithm achieves reduction in computational complexity with acceptable visual quality.

4 Conclusion

In this paper, an unvarying orthogonal search pattern with small triangle inner search method is developed for motion estimation in video applications. The experimental results show the significant speedup gain of 28.59–84.69% over OS, DS, and HS algorithm whereas maintaining related distortion in visual quality. The performance analysis clearly demonstrates the significant improvement of the developed UOSSTP

Table 4 Average speed improvement rate (SIR) for various BMAs in percentage

		Container	Akiyo	Carphone	Foreman	Miss-America	Stefan	Tennis	Salesman
UOSSTP over DS	Avg. SIR (%)	57.50	53.44	64.52	66.76	63.13	84.69	72.21	55.93
UOSSTP over OS	Avg. SIR (%)	49.64	42.23	43.54	46.30	42.51	37.62	30.96	53.39
UOSSTP over HS	Avg. SIR (%)	29.08	28.75	30.78	33.19	31.61	32.14	28.59	29.54

algorithm against existing approaches in case of computing the search points. Thus, a greater reduction in the computational complexity is attained in addition with good reconstruction quality depends upon the motion contents present on the video. The speedup gain can be further improved with some optimization algorithm. This work is part of our thesis work.

References

1. Gangodkar, D., Kumar, P., Kumar, P., Mittal, A.: Real-time motion detection using block matching algorithms on multicore processors. Int. J. Inf. Commun. Technol. (IJICT) **3**(2), 131–147 (2011)
2. Koga, T., Iinuma, K., Hirano, A., Iijima, Y., Ishiguro, T.: Motion compensated interframe coding for video conferencing. In: Proceedings of the National Telecommunications Conference, pp. G.5.3.1–G.5.3.5. New Orleans, LA (1981)
3. Li, R., Zeng, B., Liou, M.L.: A new three-step search algorithm for block motion estimation. IEEE Trans. Circ. Syst. Video Technol. **4**(4), 438–443 (1994)
4. Po, L.M., Ma, W.C.: A novel four-step search algorithm for fast block motion estimation. IEEE Trans. Circ. Syst. Video Technol. **6**(3), 313–317 (1996)
5. Zhu, S., Ma, K.K.: A new diamond search algorithm for fast block matching motion estimation. IEEE Trans. Image Process. **9**(2), 287–290 (2000)
6. Soongsathitanon, S., Woo, W.L., Dlay, S.S.: Fast search algorithms for video coding using orthogonal logarithmic search algorithm. IEEE Trans. Consum. Electron. **51**(2), 552–559 (2005)
7. Zhu, C., Lin, X., Chau, L.P.: Hexagon-based search pattern for fast block motion estimation. IEEE Trans. Circ. Syst. Video Technol. **12**(5), 349–355 (2002)
8. Kerfa, D., Belbachir, M.F.: Star diamond: an efficient algorithm for fast block matching motion estimation in H264/AVC video codec. Multimedia Tools Appl. **75**(6), 3161–3175 (2016)
9. Mukherjee, R., Biswas, B., Chakrabarti, I., Dutta, P.K., Ray, A.K.: Efficient VLSI design of adaptive rood pattern search algorithm for motion estimation of high definition videos. Microprocess. Microsyst. **45**(Part A), 105–114 (2016)
10. Shilpa, S.P., Talbar, S.N.: Fast motion estimation using modified orthogonal search algorithm for video compression. J. Signal Image Video Process. **4**(1), 123–128 (2010)
11. Jeon, G., Kim, J., Jechang, J.: Enhanced cross-diamond search algorithm for fast block motion estimation. Image Anal. Recogn. (LNCS) **4633**(2007), 481–490 (2007)
12. Purwar, R.K., Rajpal, N.: A fast block motion estimation algorithm using dynamic pattern search. J. Signal Image Video Process. **7**(1), 151–161 (2013)
13. Rajavelu, T.: Motion estimation search algorithm using new cross hexagon-diamond search pattern. Int. J. MC Square Sci. Res. **7**(1), 104–111 (2015)
14. Zhu, C., Lin, X., Chau, L.P., Po, L.M.: Enhanced hexagonal search for fast block motion estimation. IEEE Trans. Circ. Syst. Video Technol. **14**(10), 1210–1214 (2004)
15. Zhu, C., Lin, X., Chau, L.P.: Enhanced hexagonal—based search using direction—oriented inner search for motion estimation. IEEE Trans. Circ. Syst. Video Technol. **20**(1), 156–160 (2010)

Imputation of Multivariate Attribute Values in Big Data

K. Shobha and S. Nickolas

Abstract Data preprocessing plays a decisive role in achieving quality data. Massive data in real world often contain missing values and these occur mostly because of human errors and equipment errors. Several missing data handling and treating techniques exist based on the type of missing values, using statistical or data mining approaches. Efficiency in imputation can be achieved through multiple imputation techniques. But, these imputation techniques have limitations. Hence, in this paper we propose a ensemble imputation based on self-organizing competitive neural network, Adaptive Resonance Theory 2 (ART2). The imputation values are result of ensemble approach on intra-cluster non-missing value elements, which are having the shortest distance to the missing value. Goodness of the chosen imputation values is measured based on distance to other cluster elements and distance of each data values within its own clusters. The proposed algorithm also handles outliers which results in improved imputation accuracy.

1 Introduction

Data gathering, integrating, storing, and analysis of these data have become major task in data-driven world for decision-making. Data integrated from multiple heterogeneous sources like sensors, interviews, weather data, product reviews, etc., are stored and used for data mining purposes. These collected data sometimes contain faulty, corrupted or missing values, because of human error, error in data transmission, equipment malfunctioning, and misunderstanding. Usage of these faulty, corrupted, or missing data for data analysis can result in a way-out and insensible analysis [1]. Therefore, to deal with the faulty, incorrect and missing data, it is nec-

K. Shobha (✉) · S. Nickolas
National Institute of Technology,
Tiruchirappalli, Tamilnadu 620015, India
e-mail: shoeng97@gmail.com

S. Nickolas
e-mail: nickolas@nitt.edu

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_6

essary to have an effective data preprocessing technique to obtain complete, error free, and accurate data.

Handling missing value can be categorized into deletion and imputation. Deletion can be pairwise or list wise deletion. Important factor that needs to be considered in data imputation is type of missing pattern. There are three categories of missing data, (a) Missing Completely at Random (MCAR), (b) Missing at Random (MAR), and (c) Missing not at Random (MNAR) [2]. Imputation procedure involves replacing missing values with some constant, or mean values, or imputing with machine learning procedures. Machine learning based imputation method includes supervised algorithm like k-nearest neighbor, multilayer perceptron, auto-associative neural network imputation with genetic algorithms [3, 4] and with unsupervised algorithm like K-means clustering, self-organizing map (SOM) [5]. However, existing algorithms are designed to handle data pertaining to specific field. Main aim of the proposed work is to design and develop an algorithm that work with majority of data set of different fields and applications.

2 Background Study

Numerous research works have proliferated in recent years on data imputation depending on type of applications and fields. Following are some of the existing works on data imputation in different fields and applications. Broadly speaking, methods for imputation can be divided into two main categories, single imputation, and multiple imputation for different fields [6]. The imputation in meteorological data was first proposed by Cressman [7] based on interpolation. Barnes [8] extends the work of Cressman for imputation of temperature in weather data with the improvement of combining linearly the first guess field with the interpolation. Yozgatligil [9] has given a comparative study for completing missing values in meteorological data, based on spatiotemporal relation, enumerating five imputation methods. These methods include normal ratio (NR), weighted normal ratio, simple arithmetic mean, Monte Carlo Markov Chain based strategy for multiple imputation (EM-MCMC), and multilayer perceptron. They have evaluated the result using correlation dimension techniques which are dependent on spatiotemporal correlation. Elshorbagy [10] has worked in imputation of using principles of chaos theory on streaming data. Rahman [11] has proposed imputation techniques using decision trees for categorical and numerical data. Troyanskaya [12] has proposed neural network as a nonparametric concept for imputation in gene expression. Eskelson et al. [13] and Waljee et al. [14] have worked in forestry and medicine field data imputation, respectively. Imputation has a huge impact in different applications and fields and hence it plays a vital role in data preprocessing.

3 Proposed Methodology

In this paper, we propose an ensemble approach for imputing missing values in data clusters, which are formed using self-organizing competitive neural network, Adaptive Resonance Theory 2 (ART2) [15, 16].

3.1 Overview of ART2

ART2, belongs to the family of Adaptive Resonance Theory network (ART), is a second-generation network. ART2 has overcome the limitation of first generation network ART1, by supporting continuous variables, whereas ART1 supports only binary-valued input signals. ART2 architecture consists of two layers (Lay1 and Lay2) and two deciding elements (orienting subsystem and alertness parameter (ρ)). During the training process of network, data submitted to Lay1 layer undergo feature enhancement, and then are passed to Lay2 layer with bottom-up weights. These inputs of Lay2 layer turn on a racing among all outputting elements in Lay2 layer, to discover a winning entity.

The winner entity of a race is selected with highest activation value in Lay2 layer and this node is nominated as the early select of class for the featured data form. The featured data form and the prototype for initial opted class are exposed to more evaluation in the familiarizing subsystem. If the match falls within the ρ, the superior is taken as final choice, system will reach resonance state and the winning node template is updated. If the match is found to be exterior of the alertness level, then the outputting element of Lay2 layer with the afterward best activation value is elected as the preliminary match and it is subjected to the same treatment in the orienting subsystem to determine if it is an acceptable final choice or not. This progress will last until all obtainable choices attain in lessening order of the activation of the output nodes, and if none of them meets the alertness level test, then a choice of new outputting element is enrolled as a new class and its prototype is set to present input.

From the knowledge succession of ART2, it is identified that the class pattern is reorganized each time to signify the new feature pattern. With the rest of information pattern put into the ART2 network, the pattern will be retuned during the new data learning process. If the in-order patterns are arranged with minor variations, the pattern may be progressively retuned to deflect from the ‘center’ of a group. That is the pattern signify the group of input patterns well, even though there is a slight deflect in input. This is the latent advantage of ART2 and therefore marks data ordering in cluster without errors.

3.2 Different Steps in Proposed Method

Our proposed method for data imputation consists of following steps and flow diagram as shown in Fig. 1.

1. Choosing the data set and randomly inserting missing values.
2. Clustering using consensus approach.
3. Evaluating goodness of cluster.
4. Applying ensemble approach for each cluster.

Choosing the Data Set and Randomly Inserting Missing Values: Our proposed methodology is having two phase: (a) model building phase (b) evaluation phase. In model building phase, we choose data set of interest from data repository and then randomly insert missing values with variable percentage and apply Steps 2–4 from proposed method for data imputation and choose the acceptable model. This acceptable model is then applied to real-time data set to obtain complete data set without missing values which can be used for classification, prediction, and other different types of analyses.

Forming Cluster from Consensus Approach: Consensus clustering [17] is a combination of clustering information from different runs of same algorithm on the same data set or reconciling clustering information about the same data set coming from different sources, final clusters are achieved based on similarity of various clusters. This method of clustering often generates better clusters which are less sensitive to outliers and noise. Number of clusters formed using ART2 algorithm is highly dependent on ρ. Role of ρ is to determine the bottom up weights and to determine number of clusters that can be formed from given data values. Theoretically, ρ can take the value between 0 and 1, but practically the best controlling of number of clusters will happen with values in the range of 0.7–1. Hence, in our proposed method, we have used the alertness parameter ranging from 0.7 to 1, with an increment of 0.05 in each iteration. Final clusters are formed based on the similarity approach. These clusters are a combination of missing and non-missing data points. The imputation

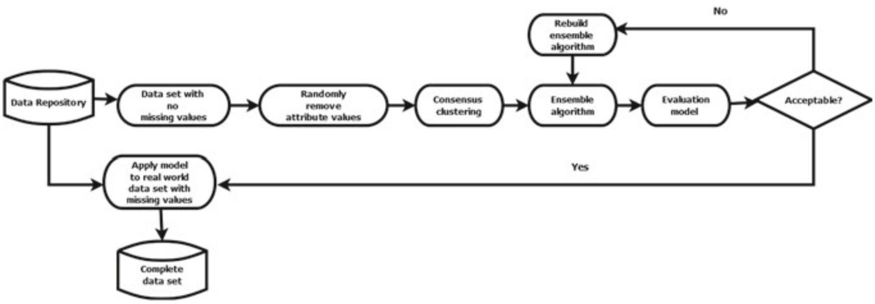

No
Rebuild ensemble algorithm
Data Repository
Data set with no missing values
Randomly remove attribute values
Consensus clustering
Ensemble algorithm
Evaluation model
Acceptable?
Yes
Apply model to real world data set with missing values
Complete data set

Fig. 1 Data flow of imputation process

error is dependent on the number of clusters and decreases if the number of clusters increases. But, too many clusters also result in accuracy drop.

Evaluating Goodness of Cluster: To assess goodness of the consensus clustering, we have used silhouette value [18]. Silhouette values are cluster evaluating techniques, which says that distance between own cluster elements should be minimum and distance to other cluster distances should be maximum.

Let C_1, $C_2 \ldots C_N$ represents clusters of data D and instance $i \in C_n$, where $n = 1, 2, \ldots N$. Let, $c(i)$ denotes the average dissimilarity of instance 'i' to all other remaining elements of C_n. Now, consider a neighboring cluster C_j, where $n \neq j$. The average dissimilarity of instance 'i' to all other remaining cluster C_j will be denoted by dis(i, C_j). After calculating dis(i, C_j) for all cluster elements of C_j, the least distance is selected, that is b(i) = min(dis(i, C_j)), $i \neq j$. This value represents the dissimilarity of instance 'i' to its nearest neighbor cluster. The silhouette coefficient S(i) is obtained from Eq. 1. where

$$S(i) = \frac{b(i) - c(i)}{\max(b(i), c(i))} \tag{1}$$

The average value S_{av} of silhouette coefficient is calculated in Eq. 2.

$$S_{av} = \sum_{i=1}^{SA} \frac{S(i)}{SA} \tag{2}$$

where SA represents sample size of cluster and S_{av} is used as evaluation criterion for cluster structure identification.

Overview of Ensemble Model: Ensemble learning is generally associated with supervised learning. Based on observations from literature study, high accuracy can be achieved by applying ensemble learning to unsupervised learning also. Upgrading imputation approach with ensemble method gives promising results with decrease in error rates which results in increased imputation accuracy. Therefore, in this proposed method, we encompass competitive neural network to an ensemble learning framework with the aim of improving the accuracy of the imputation in each cluster.

In this proposed work, ensemble approach for imputation is formed by a combination of mean of linear interpolation, mean of non-missing data points in intra-cluster, mean of entire cluster points, and by last observation carried forward methods.

4 Experimental Setup and Results

In order to evaluate the efficiency and efficacy of the proposed system, experiments were performed on a 3.50 GHz Intel(R) Xeon(R) Processor E3-1270 machine and the programming environment is Python 2.7. To check the effectiveness of proposed method experiments were conducted using ‘Pima’ diabetes data set from UCI Machine Learning Repository. This data set is a multivariate attribute of 769 records and 8 attributes with no missing values. Experiments were conducted by randomly removing data from the data set to introduce missing values in it. Missing values are created ensuring deletion from each attribute in data set, ranging from 5% to 30%, in steps of 5%. The proposed model is evaluated on Missing Completely at Random (MCAR) pattern, in which missing data are independent of each other and are random.

The accuracy of imputation is evaluated using two statistical approaches, Mean Squared Error (MSE) and Mean Absolute Error (MAE). MSE is the measure of quality of estimator giving the difference between estimators and what is estimated. MAE gives the average vertical and horizontal distance between each point.

We compared our results with mean imputation technique. This comparison exhibits that how ensemble based imputation outperformed existing mean method for “Pima” data set. Results of various missing rates of proposed algorithm and mean imputation method are shown in Table 1 and corresponding graphs in Fig. 2. From these results, one can see that the ensemble learning methods on clusters formed through consensus approach obtain less averages of MSE and MAE than the single model approach. Graphs generated from our proposed ensemble method in comparison with mean imputation method clearly explain that the proposed method is having lower error rates compared to existing mean method, which in turn proves that there is an increase in accuracy of imputation. Graph shows that error rates increase with increase in missing percentages in data set, with this we can draw the conclusion that

Table 1 Mean squared error and mean absolute error of different missing percentage

% of missing values	MAE of proposed method	MAE of mean method	MSE of proposed method	MSE of mean method
5	0.013211607	0.047347283	0.013602922	0.065388766
10	0.027886762	0.080747978	0.029334262	0.085656471
15	0.044245219	0.100500845	0.076258778	0.095300559
20	0.060784663	0.125974407	0.100460279	0.110734577
25	0.070891406	0.155897745	0.10926595	0.134589412
30	0.083196209	0.187373035	0.115723158	0.164929144

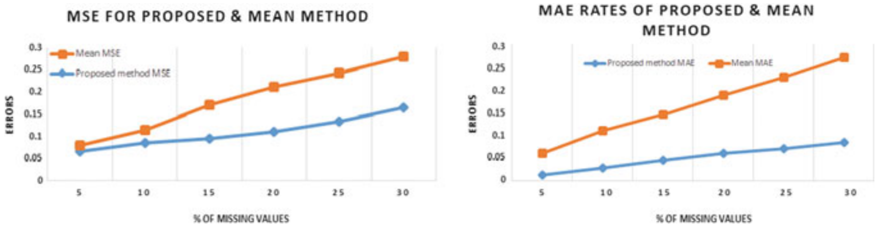


Fig. 2 Mean squared error and mean absolute error of proposed and existing method

imputation of values are worth enough when the percentage of missing is up to 25% in the data set. Greater the percentage of missing values the quality of imputation deteriorates resulting in large error rate of MSE and MAE.

5 Conclusion and Future Work

In this paper, we present a new ensemble-based imputation method, which plays vital role in the preprocessing step of data-driven world. The proposed algorithm is based on consensus unsupervised self-organizing neural network with different alertness parameters and ensemble approach. Ensemble of various imputation techniques like linear interpolation, mean imputation, and immediate last observation carried forward methods results in lower error rates in imputation compared to existing mean imputation method. Efficiency and effectiveness of the proposed algorithm are verified on benchmark data set from UCI repository. The results of experiments show that the proposed model provides lesser error rate than the conventional methods. Thus, the proposed algorithm can be used in imputation of numerical values. As a future work, proposed method will be evaluated with big data for scalability considering the impact of dimensionality reduction and dimensionality exploration.

References

1. Colantonio, A., Di Pietro, R., Ocello, A., Verde, N.V.: ABBA: adaptive bicluster-based approach to impute missing values in binary matrices. In: Proceedings of the 2010 ACM Symposium on Applied Computing, pp. 1026–1033. ACM (2010)
2. Zhu, B., He, C., Liatsis, P.: A robust missing value imputation method for noisy data. Appl. Intell. **36**(1), 61–74 (2012)
3. Gautam, C., Ravi, V.: Data imputation via evolutionary computation, clustering and a neural network. Neurocomputing **156**, 134–142 (2015)
4. Rahman, M.G., Islam, M.Z.: Missing value imputation using a fuzzy clustering-based em approach. Knowl. Inf. Syst. **46**(2), 389–422 (2016)
5. Ku, W.C., Jagadeesh, G.R., Prakash, A., Srikanthan, T.: A clustering-based approach for data-driven imputation of missing traffic data. In: 2016 IEEE Forum on Integrated and Sustainable Transportation Systems (FISTS), pp. 1–6. IEEE (2016)
6. Tutz, G., Ramzan, S.: Improved methods for the imputation of missing data by nearest neighbor methods. Comput. Stat. Data Anal. **90**, 84–99 (2015)
7. Cressman, G.P.: An operational objective analysis system. Mon. Weather Rev. **87**(10), 367–374 (1959)
8. Barnes, S.L.: A technique for maximizing details in numerical weather map analysis. J. Appl. Meteorol. **3**(4), 396–409 (1964)
9. Yozgatligil, C., Aslan, S., Iyigun, C., Batmaz, I.: Comparison of missing value imputation methods in time series: the case of turkish meteorological data. Theor. Appl. Climatol. **112**(1–2), 143–167 (2013)
10. Elshorbagy, A., Simonovic, S.P., Panu, U.S.: Estimation of missing streamflow data using principles of chaos theory. J. Hydrol. **255**(1), 123–133 (2002)

11. Rahman, G., Islam, Z.: A decision tree-based missing value imputation technique for data preprocessing. In: Proceedings of the Ninth Australasian Data Mining Conference, vol. 121, pp. 41–50. Australian Computer Society, Inc. (2011)
12. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B.: Missing value estimation methods for dna microarrays. Bioinformatics **17**(6), 520–525 (2001)
13. Eskelson, B.N.I., Hailemariam, T., LeMay, V., Barrett, T.M., Crookston, N.L., Hudak, A.T.: The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. Scand. J. For. Res. **24**(3), 235–246 (2009)
14. Waljee, A.K., Mukherjee, A., Singal, A.G., Zhang, Y., Warren, J., Balis, U., Marrero, J., Zhu, J., Higgins, P.D.R.: Comparison of imputation methods for missing laboratory data in medicine. BMJ Open **3**(8), e002847 (2013)
15. Luo, J., Chen, D.: An enhanced ART2 neural network for clustering analysis. In: First International Workshop on Knowledge Discovery and Data Mining, WKDD 2008, pp. 81–85. IEEE (2008)
16. Carpenter, G.A., Grossberg, S.: Adaptive Resonance Theory. Springer, New York (2017)
17. Goder, A., Filkov, V.: Consensus clustering algorithms: comparison and refinement. In: 2008 Proceedings of the Tenth Workshop on Algorithm Engineering and Experiments (ALENEX), pp. 109–117. SIAM (2008)
18. da Silva, S.F., Brandoli, B., Eler, D.M., Neto, J.B., Traina, A.J.M.: Silhouette-based feature selection for classification of medical images. In: 2010 IEEE 23rd International Symposium on Computer-Based Medical Systems (CBMS), pp. 315–320. IEEE (2010)

Retrofitting of Sensors in BLDC Motor Based e-Vehicle—A Step Towards Intelligent Transportation System

N. Pothirasan and M. Pallikonda Rajasekaran

Abstract Human life is the precious one which cannot be recoverable unlike other properties. Every year many life has been lost due to this road accidents. Many technologies are introduced to minimized vehicular accidents. We all know that India is developing country and there are many changes held in the path of development of our country such as change in road styles. Other than many countries India has a poor level of technologies improved. Government has introduced many modern methods but in that many or in the pathways of failure. This paper mainly focuses on road accidents occurring due to drowsy state and drunken state of drivers. Accidents can be prevented using Intelligent Transportation System (ITS) on E-Vehicle using installation of sensors of different kinds is used to detect, indicate, and prevent the road accidents. The eyeblink sensor alerts the driver in drowsy state and alarms using buzzer. To prevent the car, theft using fingerprint sensor. The main advantage of this intelligent system is to avoid population, theft, and control the road accidents. The alcohol sensor detects the alcohol from breadth and stops the engine by microcontroller. Fingerprint sensor detects the authorized persons to run the car.

1 Introduction

Increase population causes more usage of transportation system of vehicle. The increasing amount of transportation cause demand for fuel and due to this people are suffering from scarcity of fuel. The under maintenance of vehicles causes air pollution and it makes ozone hole and emission of UV rays that causes skin cancer [1, 2]. To avoid their problems, a new system of intelligent transportation establishes an electronic vehicle system called E-Vehicle [3, 4]. It was introduced to avoid speed

N. Pothirasan (✉) · M. Pallikonda Rajasekaran
Department of Electronics & Communication Engineering, Kalasalingam Academy of Research and Education, Krishnankoil 626126, Tamil Nadu, India
e-mail: gofire9988@gmail.com

M. Pallikonda Rajasekaran
e-mail: m.p.raja@klu.ac.in

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_7

and pollution [5, 6]. e-Vehicle can be used by small age people, old age people, and handicapped people. Driving is an activity that needs to done consciously with high care. One who follows rules and regulation take all safety precautions. Road accidents are the major disaster happening in India. In India, every year, thousands of people died due to road accidents. The major causes of these accidents are degradation performance, drowsiness, stressful driving, unhealthy work environments, and alcohol consumption and over speeding of drivers. There is a need for self-control system which detects driving pattern of a driver is based on the situation take necessary action to prevent further accidents. Many researchers found out that accident as a random, caused by the multifactor events resulting in unintended injury, death, and damage to property. More accidents are caused due to indiscipline an our roads, among them drunk and driving is the major factor that cause road accidents. Sleepiness in the other important factor that leads to crashes because it impairs the element of human performance that are critical to safe driving [7]. National Highway Traffic Safety Administration (NHTSA) data shows that sleepy drivers are less likely than alert drivers to take corrective actions before accidents. These problems can be solved by using modern techniques inside a car. E-vehicle has some additional features such as fingerprint sensor, alcohol sensor, and eyeblink sensor. This alternative system reduces road accident and makes a safe driving environment.

2 Previous Work

Scneel K. kommuri has proposed a technique on automatic speed tracking control of an E-vehicle using Permanent Synchronous Motor (PMSM). In this, the system of E-vehicle has a higher order sliding mode and sensor faults/failure identification module [8]. Ray Galvin has proposed a technique which develops a multivariate model and energy consumption (kwh/km), acceleration (m/s^2), power demand (w) and speed, and easily interpretable displays [9]. Pothirasan has proposed a technique on regenerative braking system using brushless DC motor to run it. Arduino microcontroller has been used to make the vehicle controller design. Forward rotating process, reverse rotating process, and stop process have all been used in the E-vehicle controller design. A three-phase inverter is the main part which has been used in alternative home appliances. E-vehicle speed measurement of no-load condition has been identified using average value has reference [10]. Stefano Rinaldi has proposed a technique in which the E-battery's performance of identification in E-vehicle has been done. Commonly, monitoring of plug-in E-vehicle's state of charging and discharging current has been identified [11]. A gateway for the communication among an E-vehicle and a local information system has been designed and deployed in the real system. H. P. Fluegal has proposed a technique in advanced simulation model and accelerated testing for the development of E-vehicle. Identification of realistic driving cycles of E-vehicle, acceleration speed, E-components test procedures, equipment, tools evaluation, and E-vehicle's performance has been made [12]. Chengwei Han has proposed a technique which has used NI-Lab VIEW Instrument to design an

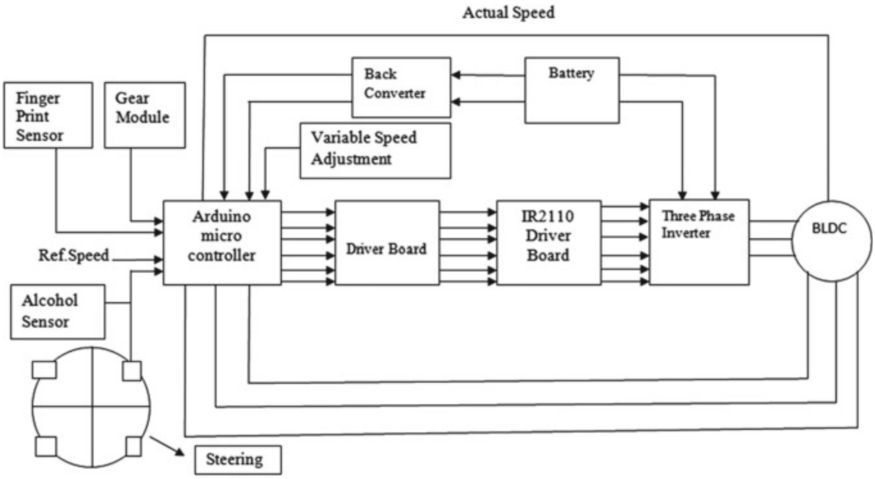


Fig. 1 Block diagram of prevention system of e-vehicle

E-vehicle physical structure, signal acquisition hardware, and subsystem E-vehicle controller. The regenerative braking system has a breaking part which has a hub motor connected to the Speed Test Target. The break experiments a result along with the motor's performance has been identified [13].

3 Prevention System of e-Vehicle

3.1 Block Diagram of Prevention System of e-Vehicle

In the proposed battery operated e-vehicle is shown in Fig. 1 the BLDC motor runs with the help of three-phase inverter. As an additional feature [14, 15], a high sensitive fingerprint sensor and an alcohol sensor was inserted. In order to produce a variation speed scheme, an accelerometer-based variable speed adjustment was used for efficient operation of Arduino microcontroller. As shown in Fig. 3 input power delivered by buck converter was used in a driver board for safety of microcontroller preventing it from voltage drop-out on the other band. The additional alcohol sensor was inserted in the four corners of the e-vehicle.

As the vehicle is designed for movement in forward and reverse direction, braking system shown in Fig. 4 was provided in the form of gear [16]. Whenever a person attempts to get into the car, the fingerprint sensor fitted asks for the fingerprint of user, after recognition of authorized finger print, the door gets unlocked, but as the ignition system of the engine is also coupled with the alcohol sensor, only when the person is alcohol free it gives ignition to the engine, thus providing additional security and safety is shown in Fig. 5.

3.2 To Examine Eyeblink Sensor in e-Vehicle

Infrared sensor is otherwise called as eyeblink sensor. It detects the blinking and analysis of blink duration of the driver's eye. As the driver's eye closes, during this process of blink detection, a vibratory sensor which has already been fixed on their seat belt startsvibrating and alerts the driver.

3.3 To Analyze Fingerprint Sensor in e-Vehicle

Fingerprint sensor is used to activate the car by an authorized person. A fingerprint sensor has attached with an E-Vehicle sensor. An unauthorized person uses this sensor the minimum 3 wrong trials.

3.4 To Proclaim Alcohol Sensor in e-Vehicle

Alcohol sensor detects the person inside the car has consumed alcohol. This project MQ3 alcohol sensor is one of a series of easy to use gas sensor that can be directly connected to an ARDUINO. This sensor measures the alcohol content in the driver by their breathing MQ3 gas sensor has high sensitivity to alcohol and good resistance towards smoke is shown in Fig. 2. This sensor is used to detect alcohol at a different concentration and it is low cost and suitable for different applications which are shown in Table 2. The value of sensor value is equal to value of sensor voltage divided by 1024. The value R0 of the sensor is equal to RS/70 and is shown in Table 1.

Table 1 Specification of alcohol sensor

Item	Parameter	Min. value	Max. value	Typical	Unit
VCC	Working voltage	4.9	5.1	5	V
PH	Heating consumption	0.5	750	–	mW
RL	Load resistance	–	–	Adjustable	
RH	Heating resistance	–	–	33	Ω
RS	Sensing resistance	1	8	–	MΩ
Scope	Detecting concentration	0.05	10	–	mg/L

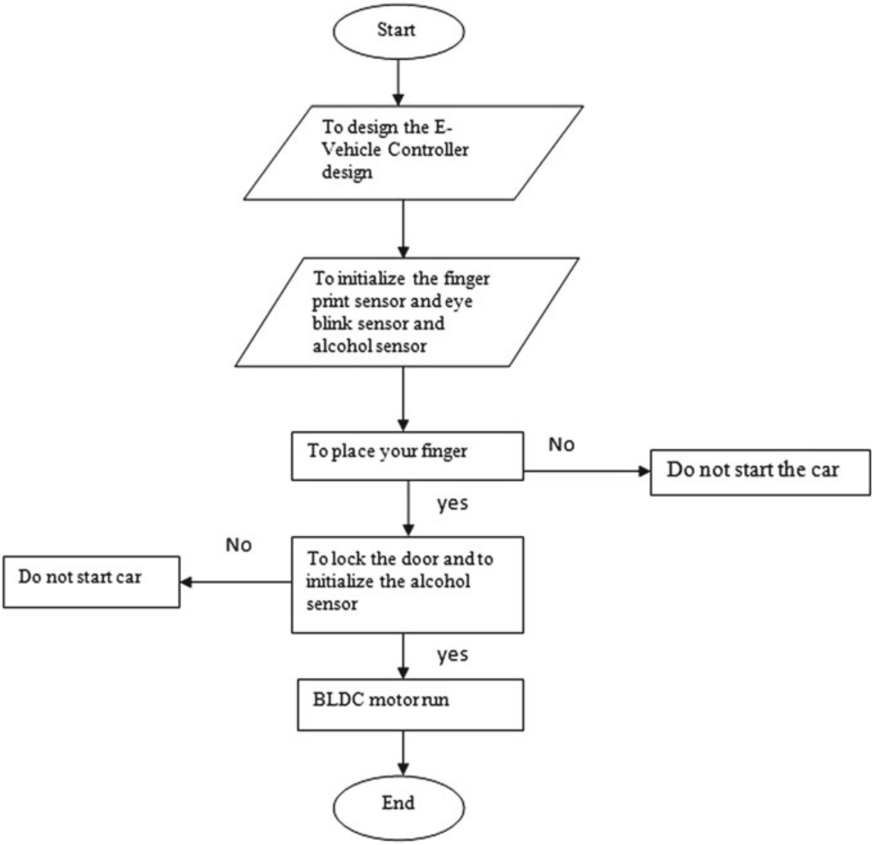

Start
To design the E-Vehicle Controller design
To initialize the finger print sensor and eye blink sensor and alcohol sensor
To place your finger
No
Do not start the car
yes
No
Do not start car
To lock the door and to initialize the alcohol sensor
yes
BLDC motor run
End

Fig. 2 Flow chart for prevention system of vehicle

Table 2 Output of the alcohol sensor reading

Sensor volt (V)	Alcohol sensor value (R0)	Alcohol condition
0.90	0.07	Present alcohol not there
1.13	0.06	Present alcohol not there
2.22	0.02	Present alcohol there
2.86	0.01	Present alcohol there

sensor volt = sensor Value/1024 * 5.0.
RS air = (5.0−sensor_volt)/sensor volt.
R0 = RS air/70.

3.5 Flow Chart for Prevention System of Vehicle

Steps 1: start
Steps 2: access to e-vehicle
Steps 3: initialize the fingerprint, alcohol, and eyeblink sensor
Steps 4: authentication using fingerprint identification
Steps 5: to unlock the door and detection of alcohol
Steps 6: ignite the engine in the absence of alcohol
Steps 7: move the car
Steps 8: stop.

4 Result and Discussion

4.1 PWM Generation of BLDC Motor

See Fig. 3.

4.2 Forward Rotation of e-Vehicle Output Waveform

See Fig. 4.

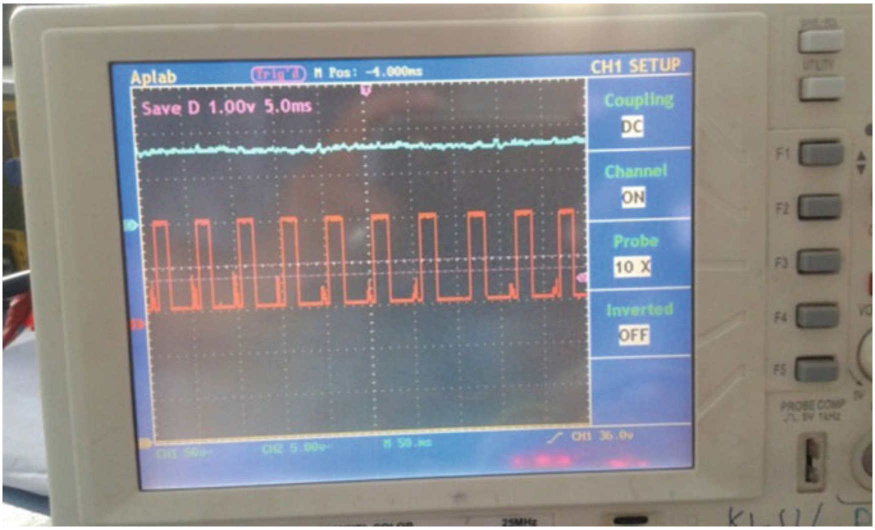


Fig. 3 PWM generation of BLDC motor

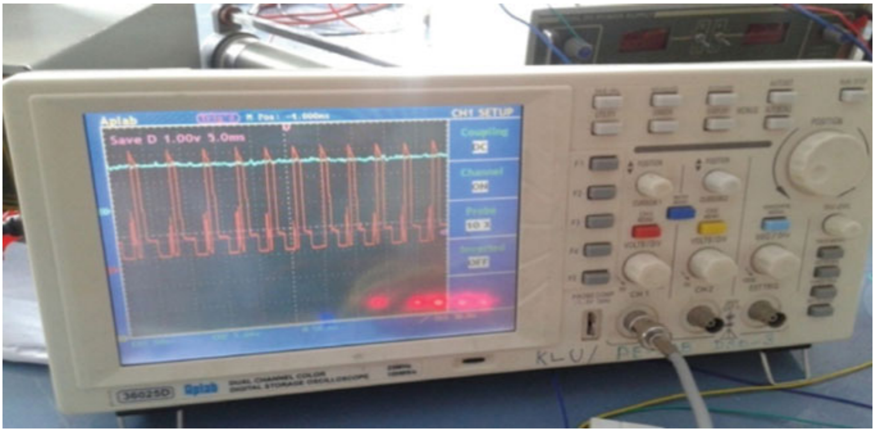


Fig. 4 Forward rotation of e-vehicle output waveform

4.3 *Over All Setup of Retrofitting of Sensors in BLDC Motor Based e-Vehicle*

See Fig. 5.

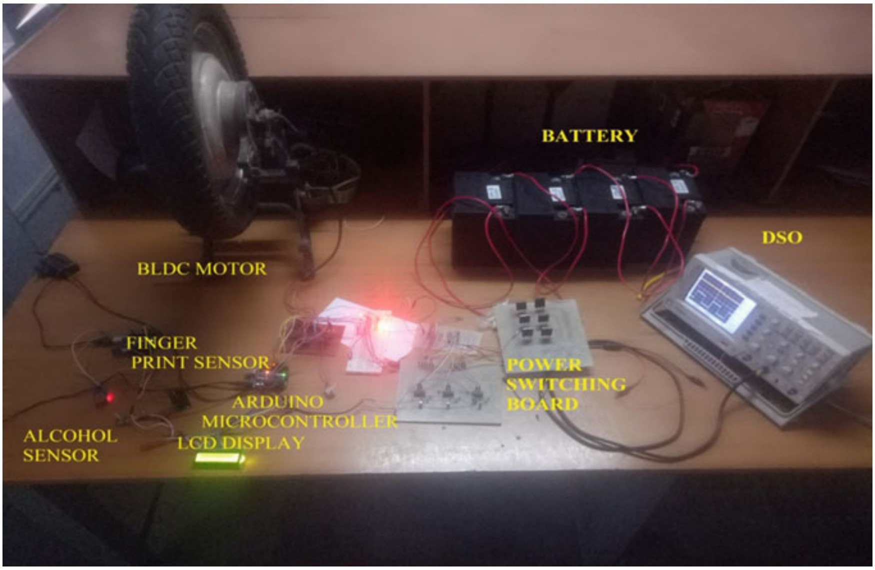


Fig. 5 Over all Setup of retrofitting of sensors in BLDC motor based e-vehicle

5 Conclusion

As the mobility rate due to car accident increases to a large extent, it is mainly due to drunken drivers and violation of traffic rules, such as crossing the speed limit. Retrofitting of various sensors such as fingerprint sensor, alcohol sensor, and speed limit sensor avoids the disadvantages of traffic rules violation and the battery operating capability of the vehicle tends to lower the pollution. In future, it is proposed to develop a solar power based e-vehicle controller design. Also, autonomous vehicle which eliminates the human errors which can be constructed using internet of things receives greater attention

Acknowledgements The authors would like to thank the Department of Electronics and Communication Engineering of Kalasalingam Academy of Research and Education, Tamil Nadu, India for permitting to use the computational facilities available in Centre for Research in Signal Processing and VLSI Design, which was set up with the support of the Department of Science and Technology (DST) Reference No: SR/FST/ETI-336/2013 dated November 2013).

References

1. Pany, P., Singh, R.K., Tripathi, R.K.: Active load current sharing in fuel cell and battery fed DC motor drive for electric vehicle application. Energy Convers. Manage. **122**, 195–206 (2016). https://doi.org/10.1016/j.enconman.2016.05.062
2. Yang, H.: Zhang, Y., Yuan, G., Paul Walker, D., Zhang, N.: Hybrid synchronized PWM schemes for closed loop current control of high power motor drives. IEEE Trans. Ind. Electron. https://doi.org/10.1109/tie.2017.2686298
3. Yang, H., Zhang, Y., Paul Walker, D., Zhang, N., Xia, B.: A method to start rotating induction motor based on speed sensorless model predictive control. IEEE Trans. Energy Convers. https://doi.org/10.1109/tec.2016.2614670
4. Walker, P., Zhu, B., Zhang, N.: Power train dynamics and control of a two speed dual clutch transmission for electric vehicles. Mech. Syst. Signal Process. **85**, 1–15 (2017). https://doi.org/10.1016/j.ymssp.2016.07.043
5. Chau, K.T., Chan, C.C., Liu, C.: Overview of permanent-magnet brushless drives for electric and hybrid electric vehicles. IEEE Trans. Ind. Electron. **55**(6) (2008). https://doi.org/10.1109/tie.2007.918403
6. Chau1, K.T., Zhang, D., Jiang, J.Z., Liu, C., Zhang, Y.: Design of a magnetic-geared outer-rotor permanent-magnet brushless motor for electric vehicles. IEEE Trans. Magn. **43**(6) (2007). https://doi.org/10.1109/tmag.2007.893714
7. Murgovski, N., Sjöberg, J.: Predictive cruise control with autonomous overtaking. In: IEEE 54th Annual Conference on Decision and Control (CDC), Dec15–18. Osaka, Japan (2015)
8. Kommuri, S.K., Defoort, M., Karimi, H.R., Veluvolu, K.C.: A robust observer-based sensor fault-tolerant control for PMSM in electric vehicles. IEEE Trans. Ind. Electron. (2016). https://doi.org/10.1109/tie.2590993
9. Galvin, R.: Energy consumption effects of speed and acceleration in electric vehicles: laboratory case studies and implications for drivers and policymakers. Transp. Res. Part D **53**, 234–248 (2017)
10. Pothirasan, N., Pallikonda Rajasekaran, M.: Regenerative e-vehicle using BLDC motor. In: International Conference on Emerging Technological Trends (ICETT) (2016)
11. Rinaldi, S., Pasetti, M., Trioni, M., Vivacqua, G.: On the integration of e-vehicle data for advanced management of private electrical charging systems. IEEE Instrum. Measur. (2017)

12. Pfluegl, H., Diwoky, F., Brunnsteiner, B., Schlemmer, E., Olofsson, Y., Groot, J., Piu, A., Magnin, R., Sellier, F., Sarrazin, M., Berzi, L., Delogu, M., Katrašnik, T., Kaufmann, A.: ASTERICS—Advanced simulation models and accelerated testing for the development of electric vehicles. In: 6th Transport Research Arena (2016). https://doi.org/10.1016/j.trpro.2016.05.432
13. Han, C., Qi, Z., Qiu, H.: Test platform design for regenerative braking of hub-motor. Han, C,. et al. (eds.) Cogent Engineering, vol. 3, p. 1253232 (2016). http://dx.doi.org/10.1080/23311916.2016.1253232
14. Nian, X., Peng, F., Zhang, H.: Regenerative braking system of electric vehicle driven by brushless DC motor. IEEE Trans. Ind. Electron. **61**(10), 5798–5808 (2014)
15. Wang, F., Yin, X., Luoline, H., Huang, Y.: A series regenerative braking control strategy based on hybrid-power. In: International Conference on Computer Distributed Control and Intelligent Environmental Monitoring (2012). https://doi.org/10.1109/cdciem.2012.22
16. Dahmani, H., Chadli, M., Rabhi, A., El Hajjaji, A.: Road curvature estimation for vehicle lane departure detection using a robust Takagi–Sugeno fuzzy observer. Int. J. Veh. Mech. Mobility **51**(5), 581–599. https://doi.org/10.1080/00423114.2011.642806

Statistical Analysis of EMG-Based Features for Different Hand Movements

C. N. Savithri and E. Priya

Abstract The electrical activity of the muscles is analyzed by surface Electromyography (sEMG). EMG signals are the essential source of control for upper limb prosthetics and orthotics and also find numerous applications in biomedical engineering and rehabilitation fields. This work focuses on the analysis of sEMG signals acquired for three different hand actions using Analysis of Variance (ANOVA) for understanding the variability of features. A single-channel sEMG amplifier is designed and signals are recorded for three different hand movements from normal subjects. Empirical Mode Decomposition (EMD) is applied to denoise the signal from artifacts. Features are extracted in time, spectral, and wavelet domain. The prominent features are selected using fuzzy entropy measure. ANOVA on prominent features shows a linear relationship between features and different hand movements and therefore these prominent features can be used to activate the prosthetic hand.

1 Introduction

Modern technology has broadened the choice of sEMG signal in clinical diagnosis, biomedical engineering, and applications [1, 2]. The characteristics of hand movement owing to muscle contraction can be obtained from sEMG signal which is a manifestation of electrical potential in time-varying form. Single-channel acquisition with surface electrodes is used to record sEMG signals as an alternative of multichannel system [3]. The random nature of sEMG signal makes it unsuitable to extract the inherent properties from solitary feature and does not permit to use these signals directly in prosthetic applications. Various sources of noise disturbing the sEMG signal are electrode noise, electrode, and cable motion artifact, power

C. N. Savithri (✉) · E. Priya
Department of Electronics and Communication Engineering, Sri Sai Ram Engineering College,
Sai Leo Nagar, West Tambaram, Chennai 600044, India
e-mail: savithri.ece@sairam.edu.in

E. Priya
e-mail: priya.ece@sairam.edu.in

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_8

line interference [4–7]. Several techniques are used for the removal of these types of noise. Filtering is a significant preprocessing technique that removes noise from the acquired signal. Few techniques include baseline wander correction accomplished by a combination of EEMD and morphological filtering [8], and canonical correlation analysis followed by morphological filtering for removal of additive white Gaussian noise [9]. The optimal window length is 150–250 ms to extract significant features from sEMG signal [10, 11]. It is imperative to extract feature vector from the input data so that it enhances the further processing stages probably a controller to actuate the prosthetic hand [12].

In this work, one-channel sEMG amplifier is developed to acquire sEMG signals for three different hand movements. Preprocessing step involves noise removal by Empirical Mode Decomposition (EMD) technique. Features in time, spectral and wavelet domain are extracted from preprocessed signal. Fuzzy entropy based feature reduction method is attempted to identify the best feature among different hand actions. The relationship between features and hand action is analyzed by Analysis of Variance (ANOVA) to get a good insight feature set and hand actions and to identify prominent features that will drive the controller of prosthetic hand.

2 Methodology

2.1 EMG Signal Acquisition

A one-channel sEMG acquisition system is designed and sEMG signal is recorded from thirteen normal subjects for three hand actions namely closed fist, spherical grasp, and point. The subjects are requested to perform one category of all hand action five times in each trial with a rest–motion–rest pattern. The muscle fatigue and mental stress to subjects are avoided by relaxing them for a minute between every hand motion.

The block diagram of one-channel sEMG amplifier developed is shown in Fig. 1. Three disposable disc surface electrodes are used one of which is a reference electrode placed over the wrist and a pair of signal electrodes on flexor digitorum superficialis muscle. Low-frequency noise and artifacts due to movements are filtered by 12 Hz RC high pass filter. A high pass filter at the second stage brings the signal to TTL level with an additional gain of 20.

The offset problems are resolved by bias adjustment that changes the reference level of the amplified signal. The sEMG signals are sampled by Analog to Digital Converter (ADC) of 10 bits resolution.

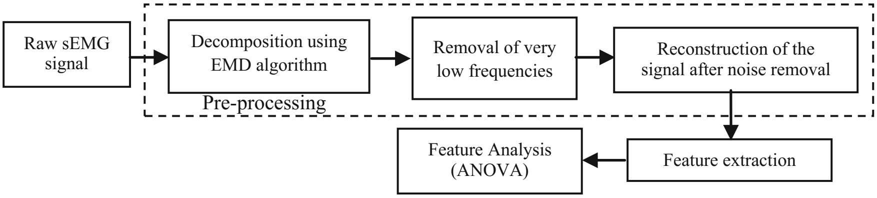

Raw sEMG signal
Decomposition using EMD algorithm
Removal of very low frequencies
Reconstruction of the signal after noise removal
Pre-processing
Feature Analysis (ANOVA)
Feature extraction

Fig. 1 Functional block diagram of sEMG signal processing

2.2 EMG Signal Preprocessing

The frequency components less than 10 Hz are eliminated by decomposition procedure and within 20–500 Hz are restored. EMD algorithm is applied as sEMG signals are nonstationary and nonlinear signals. EMD is a purely data-driven, signal-dependent procedure and makes no assumptions about the input signal [13].

EMD decompose the signal into finite number of one-dimensional function called Intrinsic Mode Functions (IMFs). Sifting algorithm performs the iterative method to find the IMFs of any given signal $x(t)$. By computing minima and maxima from the signal envelope, the mean is calculated and is subtracted from the original signal to compute the IMFs. This iterative procedure is continued until the difference remains unchanged. The whole process is repeated until $x(t)$ has more than one extremum (neither a constant nor a trend). The de-noised signal is constructed with the lower order IMFs leaving the three higher order IMFs. The time, frequency, and wavelet domain features are extracted from the reconstructed signal.

2.3 Feature Extraction

Features are distinctive properties of signal that help in differentiating between the categories of signal patterns. The mathematical formulation [14] of the features is discussed in the next few paragraphs.

The Mean or the Average Value (MAV) of the signals is obtained by averaging the absolute value of the signals over the number of signals at any time instant. MAV aid in quantifying the muscle contraction levels and is given by

$$\mathrm{MAV} = \frac{1}{N}\sum_{k=1}^{N} |x_k| \tag{1}$$

where x_k is the kth sample in the analysis among N total samples.

Integral Absolute Value (IAV) is an indication of total muscular effort and is represented by

$$\mathrm{IAV} = \sum_{k=1}^{N} |x_k| = \mathrm{MAV} \times N \tag{2}$$

The power of sEMG signals is related to non-fatiguing contraction and other forces that act on muscles. This is analyzed using Root Mean Square (RMS) value and given by

$$\mathrm{RMS} = \sqrt{\frac{1}{N}\sum_{k=1}^{N} x_k^2} \tag{3}$$

Waveform Length (WL) is the cumulative length of the EMG signal and is given by,

$$\mathrm{WL} = \sum_{k=1}^{N} |\Delta x_k| \quad \text{where } \Delta x_k = x_k - x_{k-1} \tag{4}$$

Muscle contraction state data can be obtained from Auto Regression (AR) coefficients. Generally, linear autoregressive time series are used in modeling individual EMG signals

$$x_k = \sum_{i=1}^{p} a_i x_{k-1} + e_k \tag{5}$$

where a_i presents autoregressive coefficients, p is the order of AR model, and e_k is the residual white noise. The frequency with which signal crosses zero is given by Zero Crossing (ZC), and it is linked to the original signal frequency. Low-level noise cutoff is achieved by incorporating a threshold ε.

$$\begin{aligned}&\{x_k > 0 \,\text{and}\, x_{k+1} < 0\} \,\text{or}\, \{x_k < 0 \,\text{and}\, x_{k+1} > 0\} \,\text{and}\\ &|x_k - x_{k+1}| \geq \varepsilon\end{aligned} \tag{6}$$

Signals can sometimes exceed the threshold. The frequency of the change of EMG signal is Wilson Amplitude (WA) and is indicative of muscle contraction level.

$$\mathrm{WA} = \sum_{i=1}^{N} f(|x_i - x_{i-1}|) \quad f(x) = \begin{cases} 1 \text{ if } x > \text{Threshold} \\ 0 \quad < \text{Threshold} \end{cases} \tag{7}$$

Wavelet transform finds numerous applications in bio-signal processing [15, 16]. Discrete Wavelet Transform (DWT) is performed to decompose the signal into four levels with coiflet (coif) as mother wavelet. Better performance analysis of EMG signal is obtained by decomposing the signal to four levels [17]. Relevant features are computed from the fourth level approximation coefficients.

2.4 Entropy-Based Feature Selection

Fuzzy entropy is used to express the mathematical values of the fuzziness of fuzzy sets. It is a measure of uncertainty or vagueness that exists for a given data set. The fuzzy entropy value $H(A)$ is computed with the similarity values $\mu(x_i)$.

$$H(A) = -\sum_{i=1}^{n} \mu_A(x_i)\,\log\mu_A(x_i) + (1 - \mu_A(x_i))\,\log(1 - \mu_A(x_i)) \tag{8}$$

A high similarity is indicated by low fuzzy entropy value and a feature is eliminated from feature set that has highest entropy value [18]. Prominent features are computed by iterating the above procedure.

2.5 Analysis of Variance (ANOVA)

ANOVA is a statistical tool for testing the hypothesis by measuring the variability within groups that has a baseline against which differences among group means can be compared [19]. The variability within groups (SSW) and between groups (SSB) is compared to determine if the group means are significantly different. The strength of the relationship between-group membership and the variable measured is quantified by a descriptive statistic parameter R^2. The $F -$ ratio (F) is the ratio of variance of the group means to mean of the within-group variances. It gives an insight to whether there exists a significant difference in variance between the means of two populations. The percentage variance for between-group variations is given by a parameter called critical value (f_{crit}). Thus, analysis of variance is performed to explore suitable relationship that exists between features and hand movements.

3 Results and Discussion

The sEMG amplifier picks up the raw sEMG signal for three movements with each action repeated five times. Figure 2 shows hand gestures for three different actions and Fig. 3a shows a typical waveform for closed fist. The “rest-motion-rest” movement and the offset shift in voltage are shown by single burst in Fig. 3a.

The sEMG signal is de-noised using EMD algorithm which resolves the signal into eleven IMFs. The preprocessed signal is obtained by reconstructing the signal omitting the three higher order IMFs and nullified offset as shown in Fig. 3b. The decomposition of the signal by EMD is shown in Fig. 3c. The computed features set of eight each from time; frequency and wavelet domain are shown in Table 1.

Fuzzy entropy measure selects the prominent features from the feature set presented in Table 1. Entropy values calculated after each iteration is listed in Table 1

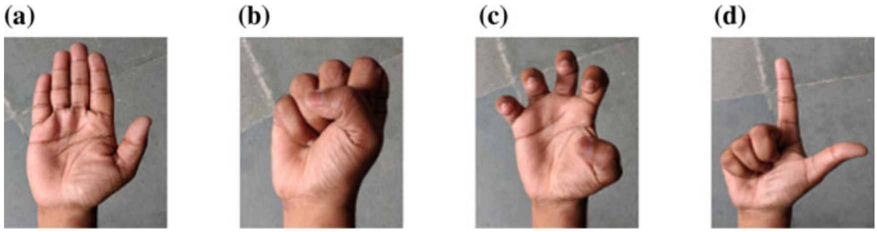


Fig. 2 Different hand gestures **a** rest, **b** closed fist, **c** spherical grasp and **d** point

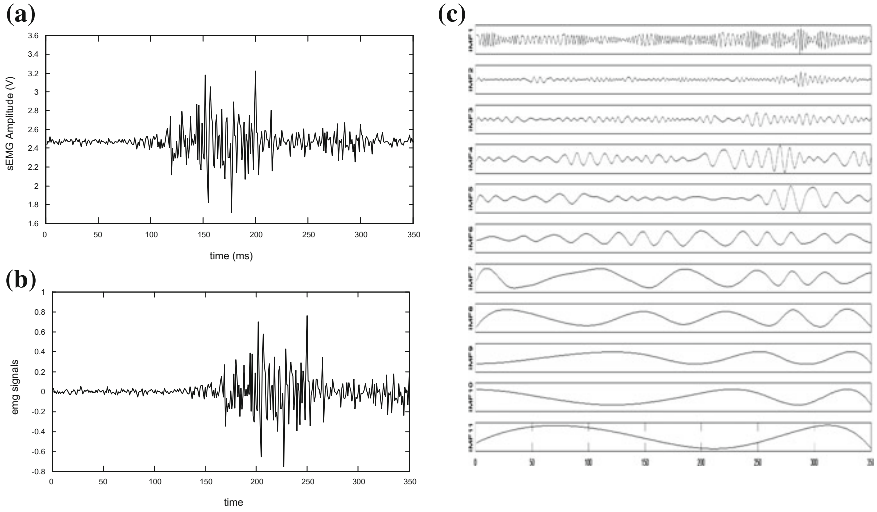


Fig. 3 Typical (**a**) raw (**b**) reconstructed sEMG signal for closed fist (**c**) corresponding IMFs of EMD

and the blank entry represents the elimination of the feature Drop in Power density ratio (DP) during first iteration. Actually, the feature with highest entropy value is removed as the impact of sEMG signal on that feature is very less and the process is repeated with remaining features. Finally, the features Zero Crossing (ZC), mean frequency (meanf) and WA qualify as prominent features in time, frequency and wavelet domain respectively which are used to control the prosthetic hand.

Relational interpretations between the features and hand actions are studied by calculating the statistical parameters of one way ANOVA. R^2 value is computed from the results of ANOVA which equals the between-group sum-of squares to total sum-of-squares. Higher value indicates a linear relationship with feature set and hand actions. Table 2 shows the R^2 values computed for all the 24 features together in time, spectral, and wavelet domain against each hand action.

The feature set provides direct measurement of force involved while performing various hand actions. It is observed that there exists a significant difference between all features across domain and hand action. In all cases, the F (9.71) is larger than

Table 1 Feature set and their corresponding fuzzy entropy measure

Time domain		Frequency domain		Wavelet domain	
Feature	Entropy value	Feature	Entropy value	Feature	Entropy value
feat_MAV	22.22	DP	–	feat_RMS	24.69
feat_RMS	17.30	**meanf**	**15.42**	feat_WL	24.41
feat_IAV	21.56	medianf	17.15	feat_ZC	23.26
feat_SSC	20.85	SM0	23.63	feat_ar1	17.99
feat_WL	19.19	SM1	23.29	feat_ar2	17.19
feat_ZC	**16.42**	SM2	23.05	feat_aac	24.21
feat_ar1	18.64	OHM	24.63	feat_aav	24.18
feat_ar2	16.75	SM	25.49	**feat_WA**	**10.98**

Table 2 R^2 values for hand actions and features across different domain

Hand actions	Time domain features	Frequency domain features	Wavelet domain features
Closed fist + Spherical grasp + Point	0.60	0.56	0.63

Table 3 R^2 values for all hand actions and all features across different domain

Hand actions	Time domain features	Frequency domain features	Wavelet domain features
Closed fist	0.727	0.575	0.737
Spherical grasp	0.676	0.676	0.731
Point	0.518	0.471	0.534

the critical value f_{crit} (1.99), which implies means are significantly different. Hence, there is significant difference between the groups (SSB) than within groups (SSW) and p-values are found to be less than 0.05. Subsequently, the null hypothesis of equal means is rejected and the test statistic is significant at this level. Thus, the relationship between feature set and hand actions are not linear.

In search of feature set with higher linear relationship, next analysis of variance is computed considering all features for three actions separately for time; spectral and wavelet domain. Table 3 shows R^2 values for three hand actions separately and all features across different domains.

It is observed again that there is a significant difference in SSB than SSW. In this case also, the F is larger than the critical value f_{crit}, which implies means are significantly different. The p-values are found to be less than 0.05. Subsequently, the null hypothesis of equal means is rejected and the test statistic is significant at this level too. Thus, the relationship between features and hand actions is not linear.

The next analysis of variance is computed for the prominent feature selected by fuzzy entropy measure for three actions in time, spectral and wavelet domain. It is

Table 4 ANOVA results for feature WA for all hand action in wavelet domain

Source of variation	SS	Df	MS	F	P-value	f_{crit}
SSB	0.0026	2	0.0013	0.0168	0.983	6.16
SSW	2.791704	36	0.077			
Total	2.794304	38				

observed from Table 1 that among ZC, meanf and WA, the most significant feature is WA as the entropy value is the least and hence ANOVA is performed for WA.

It is observed that F is lesser than critical value f_{crit} (6.16) also, the p-value is large ($p > 0.005$) for the three prominent features and hence the null hypothesis is not rejected, i.e., this feature satisfies the null hypothesis. The R^2 computed through one way ANOVA is 0.91 for ZC, 0.83 in for meanf and 0.99 for WA indicating a linear relationship between the features and hand actions. Table 4 shows ANOVA results for WA in wavelet domain and it exhibits higher linear relationship among the three significant features. This indicates that muscle contraction is linear with hand action that is being performed.

4 Conclusion

This work presents an analysis based on fuzzy entropy measure and ANOVA for the exhaustive feature set derived from the acquired de-noised sEMG signal. The features ZC in time domain, meanf in frequency domain, and WA in wavelet domain are identified as effective features based on the fuzzy entropy measure. It is observed from ANOVA results when exhaustive feature set is considered the relationship is not linear between features and hand actions. This is due to the fact that variance between groups is very less leading to small R^2 value. Hence, an exhaustive feature set cannot be used to actuate controller that drives prosthetic arm as it increases the computational load and may increase the time of response. The variability between group increases as feature dimensionality reduces. ANOVA is very promising as it reveals a linear relationship between prominent features selected by fuzzy entropy measure and the hand movements. As a future work, it is planned to design a controller using these effective features for the control action of prosthetic hand, as they show linear relationship with hand movements.

Declaration Authors have taken the consent from the concerned authority to use the materials, etc., in the paper. Authors will be solely responsible if any issues arise in future with regard to this.

References

1. Reaz, M.B., Hussain, M.S., Mohd-Yasin, F.: Techniques of EMG signal analysis: detection, processing, classification and applications. Biol. Proc. Online **8**(1), 11–35 (2006)
2. Karan, V.: Interpretation of surface electromyograms to characterize arm movement. Instrum. Sci. Technol. **42**, 513–521 (2014)
3. Phinyomark, A., Phukpattaranont, P., Limsakul, C.: Fractal analysis features for weak and single channel upper-limb EMG signals. Expert Syst. Appl. **39**, 11156–11163 (2012)
4. Baspinar, U., Varol, H.S., Senyurek, V.Y.: Performance comparison of artificial neural network and Gaussian mixture modeling classifying hand motions by using sEMG signals. Biocybern. Biomed. Eng. **33**(1), 33–45 (2013)
5. Wang, N., Chen, Y., Zhang, X.: Realtime recognition of multi-finger prehensile gestures. Biomed. Sig. Process. Control **13**, 262–269 (2014)
6. Smith, L.H., Hargrove, L.J., Lock, B.A., Kuiken, T.A.: Determining the optimal window length for pattern recognition-based myoelectric control: balancing the competing effects of classification error and controller delay. IEEE Trans. Neural Syst. Rehabil. Eng. **19**(2), 186–192 (2011)
7. Veer, K.: Experimental study and characterization of sEMG signals for upper limbs. Fluctuation Noise Lett. **14**, 150028 (2015)
8. Tiwari, D.K., Bhateja, V., Anand, D., Srivastava, A., Omar, Z.: Combination of EEMD and morphological filtering for baseline wander correction in EMG signals. In: Proceedings of 2nd International Conference on Micro-Electronics, Electromagnetics and Telecommunications, pp. 365–373, Springer, Singapore (2018)
9. Anand, D., Bhateja, V., Srivastava, A., Tiwari, D.K.,: An approach for the preprocessing of EMG signals using canonical correlation analysis. In: Smart Computing and Informatics, pp. 201–208. Springer, Singapore (2018)
10. Tsai, A.C., Hsieh, T.H., Luh, J.J., Lin, T.T.: A comparison of upper-limb motion pattern recognition using EMG signals during dynamic and isometric muscle contractions. Biomed. Sig. Process. Control **11**, 17–26 (2014)
11. Tavakolan, M., Xiao, Z.G., Menon, C.: A preliminary investigation assessing the viability of classifying hand postures in seniors. Biomed. Eng. Online **10**(1), 1 (2011)
12. Oskoei, M.A., Hu, H.: A survey-myoelectric control systems. Biomed. Sig. Process. Control **2**, 275–294 (2007)
13. Zhang, X., Zhou, P.: Filtering of surface EMG using ensemble empirical mode decomposition. Med. Eng. Phys. **35**(4), 537–542 (2013)
14. Vidya, K.V., Priya, E.: Frailty analysis of semg signals for different hand movements based on temporal and spectral approach. Biomed. Sci. Instrum. **51**, 91–98 (2015)
15. Martis, R.J., Acharya, U.R., Min, L.C.: ECG beat classification using PCA, LDA, ICA and discrete wavelet transform. Biomed. Sig. Process. Control **8**(5), 437–448 (2013)
16. Pal, S., Mitra, M.: Detection of ECG characteristic points using multiresolution wavelet analysis based selective coefficient method. Measurement **43**(2), 255–261 (2010)
17. Buranachai, C., Thanvarungkul, P., Kanatharanaa, P., Meglinski, I.: Application of wavelet analysis in optical coherence tomography for obscured pattern recognition. Laser Phys. Lett. **6**(12), 892–895 (2009)
18. Priya, E., Srinivasan, S.: Automated object and image level classification of TB images using support vector neural network classifier. Biocybern. Biomed. Eng. **36**(4), 670–678 (2016)
19. Veer, K., Sharma, T.: A novel feature extraction for robust EMG pattern recognition. J. Med. Eng. Technol. **40**(4), 149–154 (2016)

A Secure and Hybrid Approach for Key Escrow Problem and to Enhance Authentic Mobile Wallets

D. Shibin and Jaspher W. Kathrine

Abstract Mobile wallet is one of the commonly used approaches to make disbursement services under financial and business group through mobile applications in the devices such as smart phones, iPads. On the other hand, due to inadequate resource availability of the handheld devices, large-scale computing cannot be performed. When the payment is made through wireless telecommunication network, there are possibilities of threats and attacks the mobile device and the information transacted through the mobile device. In this paper, a hybrid technique is proposed to provide authentication and to solve the key escrow problem. This approach would pay a way to authorize the data and can provide a fair cryptography environment in the transaction.

1 Introduction

Mobile wallet, which is also called as mobile imbursement [1], is evolving as one of the most recurrently used application to offer imbursement services under financial regulations through mobile device and could redefine our routine with the swift reputation of mobile Internet [1]. The rapid growth in the financial service apps and the production of latest mobile devices has brought an expansion in the use of mobile payment through mobile devices. It is evident that mobile payment will become one of the familiar payment strategies in the mere future [1, 2]. On the other hand, handheld devices are resource constraint and may not be used for large computing. Nowadays, it is hard to see a store having no option to use mobile wallet such as android play, Samsung play, and apple play. Ever since the evolution of the mobile wallets, it is highly appreciated because of the reasons such as increased security in mobile wallet, retailers widely accept mobile payment, and mobile wallets can be used for online shopping, loyalty rewards programs can be stored in mobile wallet. Convenience,

D. Shibin (✉) · J. W. Kathrine
Department of Computer Sciences Technology, Karunya Institute of Technology
and Sciences, Coimbatore 641114, Tamil Nadu, India
e-mail: shibin@karunya.edu

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_9

versatility and benefit to the customers are the key factors towards the acceptance of a new technology [3]. But, since now, limited attention has been paid to improve the security of the mobile wallet. The mobile wallet app would not be accepted by the public if message authentication and privacy preserving is guaranteed for the data that is underutilization.

In mobile payment, transactions are done with mobile telecommunications network which is having huge market attractiveness, generating hot spot and growing to greater heights in community and in business field in the recent years [4]. The mobile telecommunication network acts as an intermediate between the mobile device and the computation part of the transactions. Very sensitive and confidential is transmitted over the network to make the payment for an individual using some of the existing apps such as paytm, oxigen, mobikwik, payUmoney, etc. Smartphones [4] have definite security system in place to pact with illicit access to the sensitive and confidential data [5]. The majority communal protection mechanism involved nowadays is password-based protection [4, 6]. In this password protection scheme, the user can assign password to some critical files and data [5]. But then, the drawback is that any unauthorized user may camouflage the identity to get an access into the protected data. Even patterns, biometric systems face the same problem in smart phones. The observations indicate that a strong encryption algorithm could be a best possible solution for this problem.

This paper is organized in the following ways. In the next section, the research motivation and technical background are presented. The pros and cons about the secure mobile wallet model are represented in Sect. 2. Section 3 deals with the investigation of security and its efficiency. Conclusion and future development are discussed in Sect. 4.

2 Research Motivation and Technical Background

In this section, we discuss some of the intuitions for the idea behind our model. Based on certificate less signature and dynamic rule encryption, a secure mobile wallet model has been proposed with the technicality observed from many of the ensuing references.

Digital signature is one of the superlative approaches to provide authentication and non-repudiation of the payment information when the user uses the mobile wallet [7]. Unfortunately, the digital signature discussed in the traditional public key cryptography [8] and identity-based cryptography [9] writhes from high cost of certificate management and key escrow problem [1]. The discussion from [10–14] has shown us to avoid certificate management and key escrow problems simultaneously. From this, it is observed that the design of secure mobile wallet can be done based on certificate less signature.

Trends in moving towards financial transactions using mobile payment wallet have drastically increased nowadays. Since the mobile payment applications are prone to security attacks in terms of creating vulnerabilities to the account, password, pin, and e-money, a security feature is needed to be deployed for mobile imbursement applications [3] on the Android operating system. To propose such a security scheme, dynamic rule encryption (DRE) can be implemented. DRE has the authority to guard information by encrypting the information with dynamic rules, and DRE also preserves a function as a token for authentication [3, 15].

2.1 Design Goals

With the intention of managing mobile payment security, the designed model should gratify the following requirements [1]:

- Enforceability: Enforceability is otherwise termed as unforgeability. Only the authorized users will be allowed to proceed with the transactions. None of the unauthorized can camouflage and make an unauthenticated payment or produce a fake receipt.
- Ambiguity: The real identity of the users must be kept secret and proven.
- Traceability: The business person cannot deny a customer's payment, when the customer cannot deny his/her inveterate payment. It will be an easy way for the payment service provider (PSP) to track the transaction.
- Nullifying DoS: The business person cannot rebut the payment received from the customer wherein the customer cannot deny the confirmed payment made.
- Small Overhead: Due to the resource constraint nature of mobile devices, slight overhead must be provided for both computational cost and communication overhead.

2.2 System Model

A mobile secure wallet may follow the works discussed in [16]. Based on the discussions [1–4, 16], it becomes a decision to adopt certificate less signature [17] in order to obtain unforgeability, non-repudiation, and traceability of the transmitted messages. The idea of outsourcing technique was discussed in [17] with certificate less signature in order to shrink the computation overhead [1] in mobile devices. The designed protocol may consist of the entities [1] such as customer (Jack) who would like to purchase the products sold by the merchant online, merchant (Jill) who wants to sell the products, payment service provider (PSP) who is responsible for the security and privacy [5] information regarding the payment, CSVP is an entity which shrinks the computation overhead at the client side by outsourcing the computation to the CSVP, wallet app is a mandate to perform the transaction, secure element is used to store the user credentials on the secure cloud storage, host card emulation allows the user to perform card emulation using his/her phone, NFC technology [18] indicates a set of communication protocols that enable more than one electronic devices.

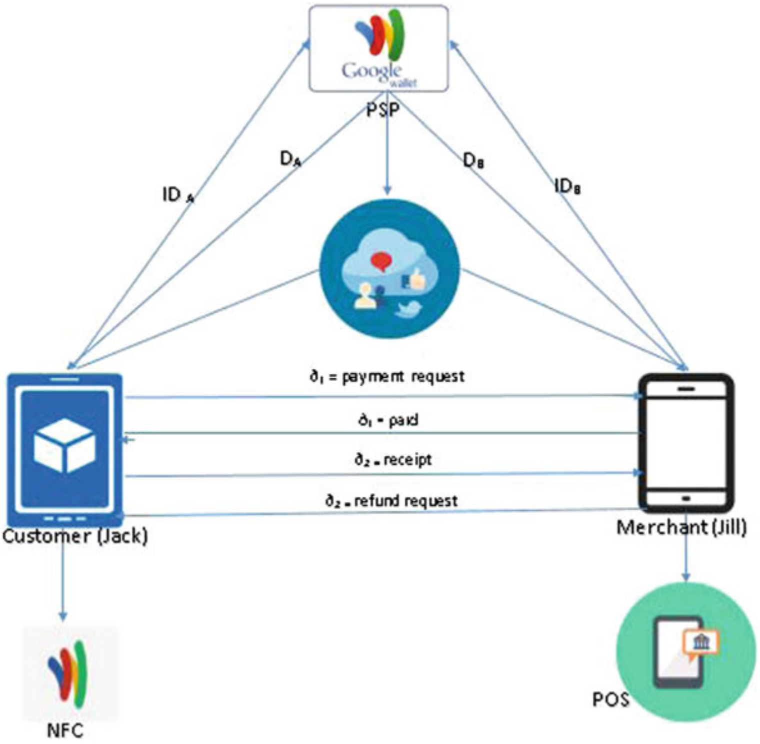


Fig. 1 System model

The interaction between the entities and the scenario is sketched in Fig. 1. The interaction [1, 2] could be classified into three categories such as setup and key generation phase, payment transaction phase, and outsourced verification phase. Mobile wallet application is considered here to provide secure transaction, less certificate cost and to solve key escrow problem.

In key setup and generation phase [1, 15], PSP inputs the input security parameter "x" and a finite description space. PSP also generates public parameters and a master key "m". It takes the client's original identity and master key as the input and produces the output as secret key $\mathrm{SK}_{\mathrm{id}}$ and a pseudo identity $P_{\mathrm{ID}.}$ There are two types of payment which include in-store payment and online payment mode. In in-store payment, Jack does online payment by using the mobile device and Jill's NFC-POS. By receiving the payment request on Jill's NFC-POS, Jack touches the NFC-POS by using his mobile phone. Jack uses his private key to sign the message using the input parameters such as transaction id, his id, and the amount transacted. The transaction is performed in a remote way using wireless connection. To achieve privacy preserving, we deploy the tamper-proof device [1, 19, 20] in order to drive pseudo-entities according to the user request. After the transaction gets completed, Jack receives an acknowledgement of reception, which will be signed by Jill's private key.

Table 1 Notations used in the model [1, 2]

Notation	Meaning	Notation	Meaning
PSP	Payment service provider	CSVP	Untrusted cloud server verification provider
POS	Point of scale	NFC	Near field communication
ID_A	Real identity of jack	ID_B	Real identity of Jill
P_{ID}	Fake identity	D	Short time partial private key
m	Master key	∂	Signature on the message

With the aim of reducing the computation overhead at the client's side, server-aided verification protocol [1, 21] can be utilized. It allows the client to transfer the received signatures to untrusted loud server verification provider CSVP wherein the CSVP performs the signature verification and produces the result to the client. The notations used in the design are given in Table 1.

As used in Google wallet [22], near field communication and host card emulation can be used to propose a secure and privacy-preserving mobile wallet.

3 Hybrid DRE and Certificate Less Signature Scheme

3.1 Proposed Model

The proposed model as sketched in Fig. 1 is carrying two stages of security concerns wherein the first phase talks about the authentication of the message and the second phase deals in solving the key escrow problem. In our scenario, the initial concern starts with the security of the message transacted from the clients end towards the merchant. The authenticity and integrity of the message need to be proven in order to have a secure transaction.

Dynamic rule encryption [2] is a symmetric block cipher with a symmetrical key [15]. Its block size is 512 bits whereas the key size is 128 bits. DRE includes six different steps [3] to be done on a plain text such as

- Pattern matrix formation
- Substitution
- Shift by rows and columns in a cyclic manner
- Transpose of matrix
- Matrix multiplication by constants
- Add round key (Fig. 2).

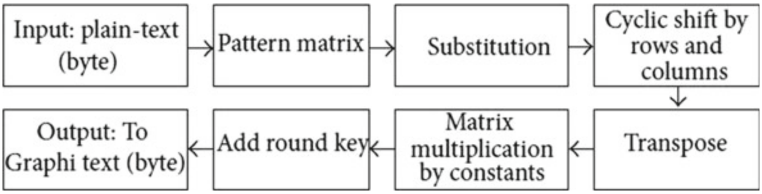
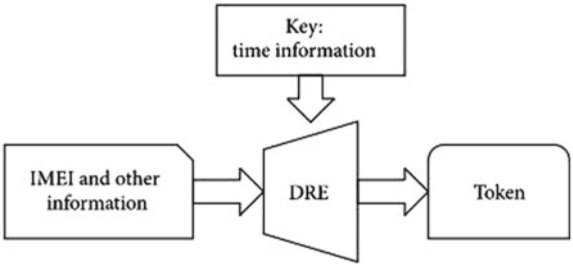

Input: plain-text (byte)
Pattern matrix
Substitution
Cyclic shift by rows and columns
Transpose
Matrix multiplication by constants
Add round key
Output: To Graphi text (byte)

Fig. 2 Stages of DRE algorithm [15]

Fig. 3 Token generator [15]



The primary motto behind DRE was to protect the data transacted during the payment using mobile apps. In mobile payment application, DRE serves as an encryption algorithm and also as a token. The role of this token is to verify that the transaction is convincing from the start till the end of the transaction. The token will be even applied to offline mode transactions since the offline mode transactions are highly susceptible to attacks. In a generated toke, the information about user's device, IMEI and IMSI numbers will also be there (Fig. 3).

The step followed during the transaction of payments is authenticated by DRE in the following ways.

a. Initially, the two devices that are going to involve in the transaction will generate the tokens.
b. The generated tokens will get exchanged between the client and the merchant. During the exchange of token, the user's information, IMEI, and IMSI numbers will be also be sent without encryption.
c. Each of the devices involved in the transaction will check the token based on the rules present in the DRE database. The merchant's token will get decrypted in the client's device so that client will obtain the information about the merchant.
d. The unencrypted IMEI and IMSI will be compared with the decrypted IMEI and IMSI numbers after exchange to check the validity of transactions.

The DRE token will get exchanged only during the transaction in order to avoid the interruption of malicious third-party devices.

Certificate less signature scheme

Certificate less signature scheme has the following stages of algorithm which includes set up phase, partial key generation phase, user key generation phase, signature, and verification phase.

In setup phase, PSP inputs the security parameter a^{sk} and returns a master secret key β, master public key P_{ub} and a list of system security parameters.

In partial key generation phase, it carries user's identity ID_{u}, params and master key β, as input and provides the user's partial private key $\mathrm{ID}_{\mathrm{u.}}$

In user key generation phase, it considers user's identity ID_{u}, params as input and precedes the user's secret/public key pair (x_{u}, Pu) after the user chooses a random value x_{u} as his secret value. Jack sends his real ID to PSP to store the credentials in the cloud storage. PSP uses tamper-proof device to generate the jack's fake identity and short time partial private key. By using his own private key, Jack computes his public key.

Payment transaction takes place as per the stages given in Fig. 1.

For signature, it takes params ID_{u}, D_{u} and X_{u} as input and provides a signature δ_i on the message m_i. For verification, it takes params $\mathrm{ID}_{\mathrm{u,}}$ the user's public key p_i and δ_i on messages $\mathrm{m_i}$ as input, then return either hold or reject.

By using the DRE encryption scheme and certificate less protocol, we can obtain unforgeability, anonymity, traceability, and non-repudiation.

3.2 Proofs Related to DRE as Tokens

a. **Time Testing**

Time testing can be defined as the weight of cipher algorithm that runs on a mobile device or an operating system [15]. Few tests have been conducted in [12–15] to measure the time factor for performing both encryption and decryption using DRE as a token. The tests have been conducted using a Google Nexus S smart phone. Also, a Sony Xperia phone was also used in this test to compare the results [2] (Table 2).

b. **Time average comparison with AES**

The test results are compared with advanced encryption algorithm [15, 23]. Both DRE and AES are used for comparison with Google Nexus device. Table 3 gives us

Table 2 Results of time testing [2]

Device	Time average of encryption (in ms)	Time average of decryption (in ms)
Google Nexus	19.537	23.519
Sony Xperia	18.262	18.428

Table 3 Time average of AES and DRE [2]

Algorithm	Time average of encryption (in ms)	Time average of decryption (in ms)
DRE	19.537	23.519
AES	0.62	1.147

the comparative results between the two algorithms. It is evident that DRE results with a fair execution time.

4 Conclusion

Mobile wallet has been creating impact in today's world by applying in all the sectors. In this paper, we have explained the reason why mobile wallet needs a great concern in terms of security and formalized the security requirements for the mobile wallet. A secure mobile wallet scheme is proposed by combining the dynamic rule encryption ad certificateless signature. To solve the key escrow problem in mobile cloud computing through the untrusted cloud server, secure outsourcing verification has been used. The real-time analysis and tracing would be the future scope of this work.

References

1. Qin, Z., Sun, J., Wahaballa, A., Zheng, W., Xiong, H., Qin, Z.: A secure and privacy-preserving mobile wallet with outsourced verification in cloud computing. Comput. Stand. Interfaces **54**, 55–60 (2017)
2. Liao, Y., He, Y., Li, F., Zhou, S.: Analysis of a mobile payment protocol with outsourced verification in cloud server and the improvement. Comput. Stand. Interfaces (2017)
3. Weiss, K.: Mobile payments, digital wallets and tunnel vision. Biometric Technol. Today **2011**, 8–9 (2011)
4. Su, H., Wen, X., Zou, D.: A secure credit recharge scheme for mobile payment system in public transport. IERI Procedia **4**, 303–308 (2013)
5. Sujithra, M., Padmavathi, G., Narayanan, S.: Mobile device data security: a cryptographic approach by outsourcing mobile data to cloud. Procedia Comput. Sci. **47**, 480–485 (2015)
6. Simoens, P., Turck, F.D., Dhoedt, B., Demeester, P.: Remote display solutions for mobile cloud computing. Computer **44**, 46–53 (2011)
7. Rivest, R.L., Shamir, A., Adleman, L.M.: A method for obtaining digital signatures and public-key cryptosystems. Commun. ACM **21**(2), 120–126 (1978)
8. Goldwasser, S., Micali, S., Rivest, R.L.: A digital signature scheme secure against adaptive chosen-message attacks. SIAM J. Comput. **17**(2), 281–308 (1988)
9. Hess, F.: Efficient identity based signature schemes based on pairings. In: Selected Areas in Cryptography, Proceedings of the 9th Annual International Workshop, SAC 2002, pp. 310–324 (2002)

10. Al-Riyami, S.S., Paterson, K.G.: Certificateless public key cryptography. In: Advances in Cryptology—ASIACRYPT 2003, Proceedings of the 9th International Conference on the Theory and Application of Cryptology and Information Security, pp. 452–473 (2003)
11. Huang, X., Susilo, W., Mu, Y., Zhang, F.: On the security of certificateless signature schemes from ASIACRYPT 2003. In: Proceedings of the 4th International Conference Cryptology and Network Security, CANS, pp. 13–25 (2005)
12. Yu, Y., et al.: Improved certificateless signature scheme provably secure in the standard model. IET Inf. Secur. **6**(2), 102 (2012)
13. Xiong, H.: Cost-effective scalable and anonymous certificateless remote authentication protocol. IEEE Trans. Inf. Forensics Secur. **9**(12), 2327–2339 (2014)
14. Xiong, H., et al.: Certificateless threshold signature secure in the standard model. Inf. Sci. **237**, 73–81 (2013)
15. Husni, E.: Dynamic rule encryption for mobile payment. Secur. Commun. Netw. **2017**, 1–11 (2017)
16. Shin, D.-H.: Towards an understanding of the consumer acceptance of mobile wallet. Comput. Hum. Behav. **25**(6), 1343–1354 (2009)
17. Huang, X., Mu, Y., Susilo, W., Wong, D.S., Wu, W.: Certificateless signatures: new schemes and security models. Comput. J. **55**(4), 457–474 (2012)
18. Coskun, V., et al.: A survey on near field communication (NFC) technology. Wirel. Pers. Commun. **71**(3), 2259–2294 (2012)
19. Zhang, C., Lu, R., Lin, X., Ho, P.H., Shen, X.: An efficient identity-based batch verification scheme for vehicular sensor networks. In: INFOCOM 2008. Proceedings of the 27th Conference on Computer Communications, pp. 816–824. IEEE (2008)
20. Kim, B.H., Choi, K.Y., Lee, J.H., Lee, D.H.: Anonymous and traceable communication using tamper-proof device for vehicular ad hoc networks. In: 2007 International Conference on Convergence Information Technology, pp. 681–686 (2007)
21. Girault, M., Lefranc, D.: Advances in cryptology—ASIACRYPT 2005. In: Proceedings of the 11th International Conference on the Theory and Application of Cryptology and Information Security, Chennai, India, Dec 4–8, Proceedings. Springer, Berlin, Heidelberg (2005), Ch. Server-Aided Verification: Theory and Practice, pp. 605–623
22. Google, Google wallet. https://www.google.com/wallet/
23. Lara-Niño, C., Miguel, A.M.-S., Arturo, D.-P.: An evaluation of AES and present ciphers for lightweight cryptography on Smartphones. In: Proceedings of the 26th International Conference on Electronics, Communications and Computers (CONIELECOMP'16), pp. 87–93. Cholula, Mexico, Feb 2016

Extended Lifetime and Reliable Data Transmission in Wireless Sensor Networks with Multiple Sinks

Vasavi Junapudi and Siba K. Udgata

Abstract In a wireless sensor network (WSN), one of the most important constraints on sensor nodes is the low power consumption requirement. Due to this, the WSN comes with a trade-off to extend the network lifetime at the cost of lower throughput or higher transmission delay. Due to small memory, the size of the buffer limits the number of data packets can hold which implies increased overflow of data packets. This congestion and collision raise questions about the reliability of the network to forward data packets. Thus, for successful monitoring of the region of interest, the data collected by the sensor nodes need to be delivered to the sink, minimizing data loss. We tried to handle the reliability problem by inducing trust model in Cluster Algorithm for Sink Selection (CASS) approach to get an assurance of reliability with low maintenance cost to enhance the network lifetime. Experimental simulations show that with our trust model incorporated in the CASS algorithm can minimize the packet drop rate in the presence of a different number of unreliable nodes while trying to maximize the lifetime of the sensor network with the use of CASS-Reliable and multiple sinks. It is also observed that the network lifetime measured in terms of the number of messages delivered successfully remains almost same in comparison to CASS in the presence of unreliable nodes.

1 Introduction

One of the most important constraints on sensor nodes is the low power consumption requirement as power sources cannot recharge or replaceable easily. Thus, a WSN comes with a trade-off to extend the network lifetime at the cost of lower throughput or higher transmission delay [1].

V. Junapudi (✉) · S. K. Udgata
School of Computers & Information Sciences,
University of Hyderabad, Hyderabad 500046, India
e-mail: vasavi.singh@gmail.com

S. K. Udgata
e-mail: skudgata@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_10

In simple, to extend network lifetime, we have two possibilities. One is increasing density of the network and second is deploying multiple sinks to receive the collected data in the target region. Now the problem can be viewed in terms of collisions and congestion due to its limitations (dense deployment, small memory to keep the data and channel for communication). Mainly, this congestion occurs in sink zone as it has to collect the data from sensors and neighbor nodes of the sink will act as a relay node for all sources to forward the received data packets to sink. Packet loss happens due to transmission errors, packet collision, interference, node failure (due to energy depletion), and buffer size in addition to congestion. Thus, congestion and collision effect the reliability of the network.

Due to multi-hopping, relay nodes' energy gets drained quickly as they participate more in communication to forward data and decreases the network lifetime. If multiple sinks get deployed, and sensor nodes can send data to any of the sinks which are nearer to it, thus saving energy of relay nodes and enhancing the network lifetime [2–4]. But deploying multiple sinks increases maintenance cost. For the reduction of increased maintenance cost due to multiple sinks deployment, CASS algorithm came with a proposal to activate the sinks in shifts means enabling a single sink at a time [5]. Once the shift rate of the sink node is below a threshold value, then CASS algorithm will choose another sink to be activated based on the network reach value.

The original CASS algorithm does not consider the presence of unreliable nodes in the network which fails to forward the data. So, in the presence of unreliable nodes, packet loss happens due to increased packet drop by unreliable nodes. Thus, there is a need for modified CASS algorithm to ensure reliable packet delivery in presence of untrustworthy nodes in the network. In this work, we propose CASS-Reliable algorithm as an extension of CASS algorithm to avoid packet loss in the presence of unreliable nodes and still achieve maximum network lifetime.

2 Related Work

Retransmission and redundancy are the major techniques used in many protocols to recover lost packets. These techniques ensure reliability by recovering the lost data using either hop-by-hop or an end-to-end method based on the end user requirement. In hop-by-hop, every hop node is responsible for reliable transmission of the data packet. In end-to-end transmission, the source and destination are responsible for maintaining the reliability, while the intermediate nodes merely relay the packets between the source and the destination [6].

In our work, we adopted iACK to know the deliverance of the data packet. We took advantage of broadcasting nature of sensor if data packet loss due to any reason the sender node will reroute the packet using another path through other neighbor nodes as data is automatically received by the neighbors also. Along with iACK, we added the trust model to strengthen the reliability concept in our work.

Trust is the level or degree of the confidence that a node can have on another node. The WSN is the collaboration of distributed sensor node to monitor the target region which can affect the security, reliability, privacy, robustness, authentication, and authorization. To deal with such issues, WSN needs a trust framework [7].

The trust value of the nodes ranges from -1 to $+1$. -1 denotes completely untrustworthy, 0 denotes moderate/acceptable trust, and $+1$ denotes completely trustworthy. In this paper, we consider only direct trust(observed by an immediate neighbor).

3 System Model

3.1 Assumptions

(a) Nodes are deployed randomly ensuring all sensor nodes will be able to connect to at least one sink.
(b) Sensors have limited energy, and initial energy is same for all.
(c) Sinks have infinite energy and can communicate with the outer world.
(d) Sensors will come to know its neighbor's information based on the distance between the node using Euclidean distance.
(e) All nodes have built-in capability to support iACK technique.
(f) Trust of all nodes initialized to zero.

3.2 Network Model

The network is modeled as a graph $G(V, E)$, where V is the set of all sensor nodes say N and the base stations/sink nodes S, i.e., $V = N \cup S$. An edge $(u, v) \in E$ if and only if nodes u and v are within the communication range. Considering sensor (referring to source) say node v is not in the communication range to reach the sink. Then source will follow the multi-hop technique to reach the sink. With multi-hop technique, the source v will find multiple paths through each neighbor to reach the sink say $P_1, P_2, ..., P_k$, where k is a positive integer and $k \geq 1$. Among all possible paths, the source node will consider the shortest path (in a number of hops) say P_j where $1 \leq j \leq k$ for minimal usage of energy due to energy constraint (discussed in Sect. 1) and the chosen path is referring as original path. Like this, a majority of the sources will have multiple paths to reach the sink.

3.3 Energy Model

An energy model on radio characteristics including energy dissipation in transmitting and receiving models to send a k-bit message to a distance d using the radio model is described in [8].

In our simulations, we assumed that

1. The radio channel is symmetric, i.e., the energy consumption to transmit the data from node A to node B is same as from node B to node A.
2. The nodes in the network would send the sensed data periodically to the sink [1].

3.4 Trust Model

We include trust model in CASS to measure reliability while extending the network lifetime by rerouting messages through other neighborhood paths. Considering the source, A has neighbors $n_1, n_2, ...n_k$ where $n_i \in V$. Source node A has chosen shortest path routed through neighbor say n_2 based on the number of hops. Assume the path is

$$P_1 : A \rightarrow n_2 \rightarrow p_1 \rightarrow p_2 \rightarrow p_3 \rightarrow Sink$$

Before sending the data to n_2, source A will check trust value of next hop (here it is n_2). If $trust(n_2) \geq 0$ then a communication link will be established, and the data packet will transmit to n_2. This process will be continued with other hop nodes to achieve reliability of the network. If any of the hop node trust value is less than zero, then the sender will look for the shortest path among all other possible neighborhood paths and update the path information in its routing table. This process continues until the sink receives the data packet.

By rerouting the data packet, we are avoiding congestion in the network, saving energy of all nodes which involved in forwarding the data and availing the residual energy of other nodes which participate less frequently in routing. By including trust model in CASS, we are achieving reliability of the network while extending the network lifetime. Of course, due to rerouting the packet, latency time has been increased with respect to an increased number of hops.

Network lifetime: As far as the node can send data and it is delivered to sink, we can consider that the network is alive to get/collect the sensed data. For our scenario, the total number of messages received by the sink is the suitable metric to define network lifetime.

Dropped messages: A dropped message is defined as the number of messages that could not be delivered to the sink in the occurrence of unreliable or untrustworthy nodes. Instead of dropping a message and retransmitting it, we are rerouting the packet through possible paths using other neighbors. This metric is used to measure the reliability of the network.

Untrusted node: A node S_j is called as untrusted node with respect to S_i where S_i and S_j are within the communication range, if the message transmitted by S_i is dropped by S_j due to some reasons then the $trust(S_i, S_j)$ will gradually decrease and makes S_j as untrusted node with respect to S_i.

Unreliable node: A node S_j is called an unreliable node, if the message transmitted by S_i is not forwarded or received by S_j due to unexpected reasons (energy depletion, buffer overflow due to small memory in size, heavy traffic at the channel) then the node S_j will be considered as an unreliable node.

4 Proposed Algorithms

Here, we are describing the method we experimented to achieve the reliability by incorporating trust model in the algorithm *Cluster algorithm for sink selection—CASS*. We are presenting the entire work in three modules. The first module explains about the initial setup of the network and the working of it using CASS [1].

Algorithm 1 Cluster Algorithm for Sink Selection - CASS

```
Input: neigh[][], path[]
Output: next active sink $S_i$
Construct tree considering sink as root
Find the nodes with at least 2 child nodes
Merge clusters whose parent is a child of another cluster
Find the pathways from one cluster to another cluster
//If one cluster has a sensor that contains a child, neighbor or parent that is part of the comparing cluster then it is one qualifying path
if $pathways \geq 2$ then
    merge cluster to increase the connectivity of the network under the sink $S_i$
end if
//Calculate network reach of sink S
network reach= no. of nodes in cluster reachable by S * i + no. of child nodes of S * j + no. of single nodes (not in cluster) reachable by S * k
//i, j, k are arbitrary weights, $\{i, j, k\} \geq 1$ & $i > j > k$
Find the sink say $S_j$ with the highest network reach
return $S_j$
```

CASS module explains the procedure to select the sink with the highest network reach to enhance the network lifetime while controlling the maintenance cost of multiple sinks and is explained in Algorithm 1. Finally, trust model is the module used to achieve the reliability of the network for message delivery by checking the trust value of each node before communication establishes and is detailed in Algorithm 2. CASS is the algorithm used to enhance the network lifetime using multiple sinks by keeping the maintenance cost under control by activating a single sink at a time (looks like a network with single sink all the time). CASS algorithm used to select the sink based on network reach. It is calculated with the help of clusters formed while constructing a tree considering sink as root and it is explained in Algorithm 1. Shift rate is the value used to know whether the sink shift is needed or not and is defined as a certain percentage of the network lost (after considerable dead nodes raised) and it does not gain the network in a certain amount of time then sink checks whether it meets the sink shift ratio or not. If shift sink network ration (reachable nodes/starting nodes) is not meeting the shift rate then continues to be active and receives the data messages otherwise CASS will be called to find the next active sink.

By using an iACK technique, we have not burdened the network with over traffic and by rerouting the data in the presence of unreliable or untrusted node, we avoided retransmitting the data packet. By doing this, we extended the lifetime of the sensor node. We calculated the message drop as one of the measures for reliability after including the trust model in CASS.

Algorithm 2 Trust model

```
Input: trust[V][V], neigh[][], path[], Energy required to act as a Source $E_T$, Energy required to
   act as Intermediate Node $E_{TR}$, source
Output: Message delivery[boolean]
   Initialization
 1: trust[][]=0
 2: nodes sends the data periodically
   //Reliable transmission of data packet
 3: repeat
 4:   Initiate data transmission using original path
 5:   sender $h_i$ checks for trust value of next hop node say $h_j$
 6:   if $trust(h_i, h_j) \geq 0$ then
 7:     data transmission will be done between $h_i$ and $h_j$
 8:     if $h_j$ is not acting as a reliable node then
 9:       reduce the value of $trust(h_i, h_j)$
10:       Find the other possible path from $h_i$ to sink through other neighbors
11:       choose the optimal path from possible paths
12:       update the original path considering the new optimal path
13:       go to Step 3
14:     else
15:       reduce the energy $E(h_i)$ and $E(h_j)$ by $E_T$ and $E_{TR}$ for sending and trans-receiving the
          data respectively
16:       increment the trust value of $trust(h_i, h_j)$
17:     end if
18:   else
19:     Find the other possible path from $h_i$ to sink through other neighbors
20:     choose the optimal path from possible paths
21:     update the original path considering the new optimal path
22:     go to step 3
23:   end if
24: until the data packet received by the sink
25: return $S_i$
```

5 Experimental Setup and Results

The proposed model is simulated on MAT LAB by deploying all nodes randomly in 100 X 100 grid. The simulation starts with the assumption that all nodes in the network (sensors and sink) have trust value zero (moderate trust). Sinks nodes are trustworthy throughout the network lifetime. We experimented our approach on 40 data sets by increasing density of sensors along with increasing the number of sinks. The experimental simulation parameters are shown in Table 1.

We experimented with three types of data sets having 100 sensors-4 sinks, 200 sensors-6 sinks, and 300 sensors-8 sinks. The number of unreliability nodes is in the range of 5–25%. We experimented the trust model in CASS for reliability considering a different number of unreliable nodes. The reliability measures in terms of dropping messages after including the trust model to assure the reliability in the presence of unreliable nodes. Results have shown in "Figs. 1 and 2" varying the value of the unreliable node from 5 to 25%. Sensor nodes will act as unreliable only for 10% of the time in its lifetime and remaining 90% will act as reliable and implemented by generating a random number to define whether the node is acting as reliable or not. The trust value is changed depending on the number of data packets forwarded or not forwarded by the node. The trust value is changing with a value 0.0001.

Table 1 Input data information

Grid size	100 × 100
Initial energy of sensors	10 J
Number of sensors	100, 200, 300
Number of sinks	4, 6, 8
Communication range	15 units
Shift rate	90%
Trust_Value (S_1, S_2)	0 where S_1 and S_2 are within the communication range
Unreliable nodes	5– 25% on total deployed sensor nodes

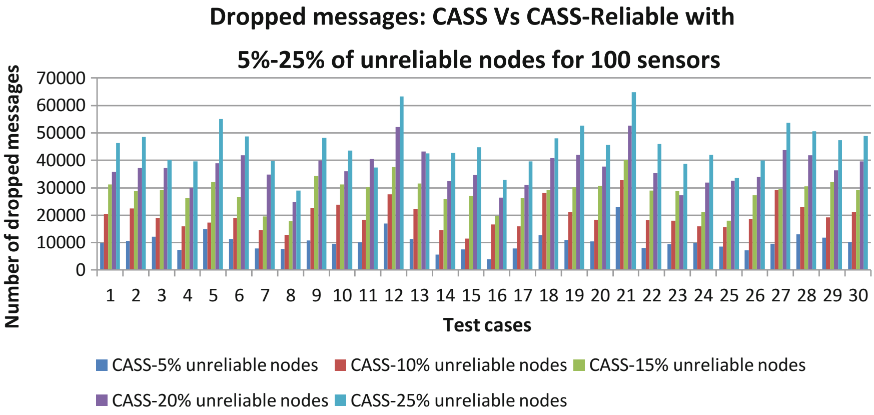


Fig. 1 Number of dropped messages in CASS model with 100 sensors

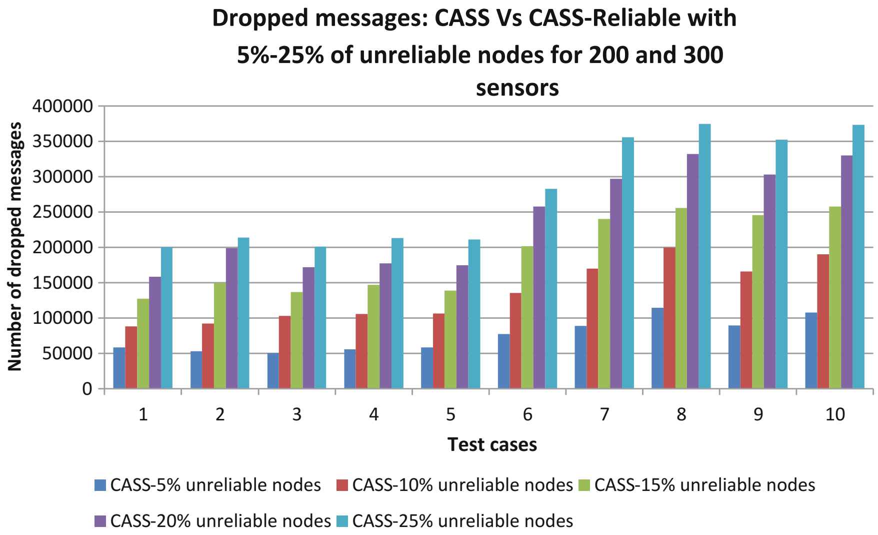


Fig. 2 Number of dropped messages in CASS model with 200 and 300 sensors

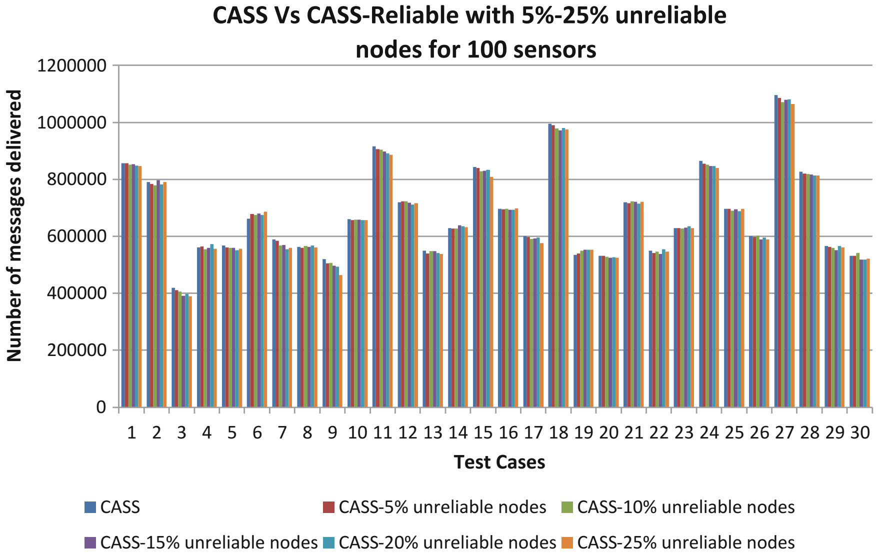


Fig. 3 Network lifetime comparison CASS Vs CASS—Reliable model with 100 sensors varying unreliable nodes

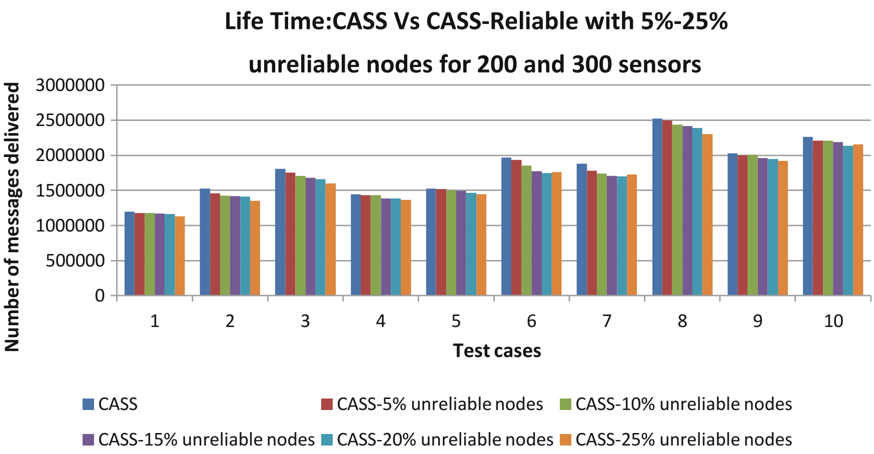


Fig. 4 Network lifetime comparison CASS Vs CASS—Reliable model with 200 and 300 sensors varying unreliable nodes

From the "Figs. 1 and 2" as expected the number of dropped messages are increasing with respect to increasing unreliable nodes. Instead of retransmitting the entire packet we rerouted the data packet from an untrusted node or an unreliable node. By doing this, we extended the network lifetime by rerouting the messages through other neighbors. By doing this, the network lifetime increased by few more messages by availing the residual energy of those nodes which are not involved in routing frequently.

We compared the network lifetime of CASS (assuming that all nodes are reliable and data sent by the source will be received by the sink for sure) with CASS-Reliable model and results are shown in "Figs. 3 and 4". Here, we can notice that even with the increased number of unreliable nodes, the sink can receive an almost same number of messages as we used alternate neighborhood path to deliver a packet.

6 Conclusion

This paper describes how best we can achieve the reliability with iACK technique and taking advantage of broadcasting nature of sensor nodes to deliver the packet with reliability without retransmitting the data. It is achieved by rerouting the data packet through other shortest neighborhood path which ensures the reliability of the defined trust model. By doing this, we can use the residual energy of other nodes to act as relay nodes in delivering the data packet. When we experimented the model on a different data set with increased sensor node density and number of sinks, it has shown considerable improvement in delivering the message in the presence of unreliable nodes. The rerouting mechanism also helped to avoid congestion and collision in the network by avoiding retransmission of packets.

References

1. Junapudi, V., Udgata, S.K.: Lifetime maximization of wireless sensor networks with multiple sinks using multiple paths and variable communication range. Int. J. Sens. Netw. https://doi.org/10.1504/IJSNET.2016.10001453
2. Chen, S., Coolbeth, M., Dinh, H., Kim, Y.-A., Wang, B.: Data collection with multiple sinks in wireless sensor networks. Wireless Algorithms, Systems and Applications, vol. 5682. Springer, Berlin Heidelberg. Lecture Notes in Computer Science, pp. 284–294 (2009). ISBN 978-3-642-03416-9
3. Das, A., Dutta, D.: Data acquisition in multiple-sink sensor networks. SIGMOBILE Mob. Comput. Commun. Rev. **9**(3), 82–85 (2005). ISSN 1559-1662
4. Vincze, Z., Vida, R., Vidacs, A.: Deploying multiple sinks in multi-hop wireless sensor networks. In: IEEE International Conference on Pervasive Services, pp.55–63. Istanbul, Turkey (2007)
5. Boler, C., Yenduri, S., Ding, W., Perkins, L., Harris, J.: To shift or not to shift: maximizing the efficiency of a wireless sensor network using multiple sinks. Int. J. Netw. Comput. Adv. Inf. Manage. (IJNCM) **1**(1), 4 (2011). https://doi.org/10.4156/IJNCM
6. Mahmood, M.A., Seah, W.K.G., Welch, I.: Reliability in wireless sensor networks: a survey and challenges ahead. Compu. Netw. **79**(Supplement C), 166–187 (2015). ISSN 1389–1286, https://doi.org/10.1016/j.comnet.2014.12.016
7. Karthik, N., Sarma Dhulipala, V.R.: Trust calculation in wireless sensor networks. Proceedings of IEEE international Conference on Electronics Computer Technology **4**, 376–380 (2011)
8. Heinzelman, W.R., Chandrakasan, A., Balakrishnan, H.: Energy-efficient communication protocol for wireless micro sensor networks. Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, p. 10. IEEE, Hawaii, USA (2000)

Prediction of Software Cost Estimation Using Spiking Neural Networks

V. Venkataiah, Ramakanta Mohanty and M. Nagaratna

Abstract In the modern competitive business world, any software firm has a success, to retain sustainability, the significant task is to estimate accurate process cost, i.e. completion time, resources, and required budget at an early stage of software development which gives success and sustainability of the industry, due to the competitive world, the project managers are estimating resources to hurry up, consequently low-quality precision. Eventually, companies are lacking in terms of metrics such as quality product, projects are completed within time, and budget. The development of the software scope is increasing in demand, the complexity in the development process also increasing. As a result, the inaccurate software product is delivered. Hence, there would be an enhanced model which estimates better planning, efficient manpower, and better resource allocation are vital in the early stages of the software development lifecycle. Therefore, the prediction of an estimated process cost of the project is highly desirable in the early stages of the software project. In this research article, the proposed model uses Spiking Neural Networks is to improve the accuracy of estimated process cost which is improving the quality of the software. There are three data sets which are used to validate the performance of the proposed technique using RMSE, MMRE statistic metrics.

V. Venkataiah (✉)
Department of CSE, CMR College of Engineering & Technology, Medchal 500055,
Hyderabad, India
e-mail: venkat.vaadaala@gmail.com

R. Mohanty
CSE Department, Keshav Memorial Institute of Technology, Narayanaguda 500011,
Hyderabad, India
e-mail: ramakanta5a@gmail.com

M. Nagaratna
CSE Department, JNTUH College of Engineering, Kukatpally 500085,
Hyderabad, India
e-mail: mratnajntu@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_11

1 Introduction

In the modern competitive business world, the companies are becoming the success, to retain sustainability, the significant task is to estimate accurate process cost, i.e. completion time, resources, and required budget at an early stage of software development which gives success and sustainability of the industry, due to competitive world and not able to finish within stipulated time, the project managers are estimating resources with lowquality precision. Eventually, companies are lacking in terms of metrics such as quality product, projects are completed within time, and budget. The development of the software scope is increasing in demand, the complexity in the development process also increasing. As a result, the inaccurate software product is delivered. Hence, there would be an enhanced model which estimates better planning, reliable design, efficient manpower, and better resource allocation are vital in the early stages of the software development lifecycle. Therefore, the prediction of an estimated cost of the process is highly desirable in the early stages of the software product development, to improve the accuracy of the process cost estimation is a biggest challenging task in computer science and software engineering. The software estimation tool became to prevent or reduce failure of the projects, and improve the quality of the product. The development of SCE is sensitive, complex and inevitable [1]. SCE is a process of predicting an amount of effort is required to complete the project is measured by means of the Person-Month (PM). Once the effort is estimated, we can estimate the estimated cost and time of the project. Subsequently, one decade, research was focused on to fill the gap between the estimated time and actual time. Finally, 90% projects have succeeded, till authors are facing a problem is the quality of the product is low. There would be highly computational learning methods to predict better-estimated process cost to maintain high quality of the product which is satisfying the need of the customers. There are numbers of predicting techniques have been proposed classified into three, such as Parametric, Non-algorithmic, and Machine Learning. Neural Network (NN) is a machine learning technique is able to compute, processing, classifying, and predicting the new kind of pattern and mapping input–output pattern. It has the ability to solve the nonlinearity kind of problems, to find the fault tolerance [2]. NN technique has a special feature to make available train knowledge for taking any kind of decision.

The main objective of this paper is to propose a methodology which is third-generation Neural Networks are also known as Spiking Neural Networks (SNNs). This technique has proved that, in many, cases resolving complex information processing tasks. Hence, SNNs are superior to traditional NNs [3, 4, 5]. This model has the ability to encode temporal data generated by use of trains of spike, allowing authors for fast decoding and multiplexing information [6]. There are two versions (1) evolving SNN (eSNN) [7] and (2) dynamic eSNN (deSNN) [8, 9]. In modern days, most widespread new computational architecture of SNN is NueCube [10]. Basically, it was designed for processing the collection of spatiotemporal brain data. At this moment, to adapt and make use in predicting modeling of time series of data for early-stage software development of the cost estimation. This technique has nowhere, nobody employed in the field of SCE till today. The rest of the paper can be organized as the following manner. In Sect. 2 Literature Review. In Sect. 3 Spiking Neural Networks. Research Methodology in Sect. 4. In Sect. 5 Results and discussions. Finally, in Sect. 6 conclusion of the paper.

2 Literature Review

In modern years, Neural Networks have been used for solving complex problems in different domains, especially in cognitive science, software engineering for prediction unknown objects. In this paper, presenting the work to predict the estimated cost and accuracy in software quality at the early stages of the software development lifecycle. Attarzadeh et al. [11] proposed a novel Artificial Neural Networks that are integrated with COCOMO II to improve the vagueness of attributes of the software product. The proposed model showed 8.36% improvement in the prediction of estimation accuracy comparable to the COCOMO II model employed on datasets are COCOMO I and NASA 93. Ghose et al. [12] presented a comparison result of various Neural Networks techniques are, viz. FFBPNN, Cascaded FFBPNN, Elman BPNN, Layer Recurrent NN, and Generalized Regression NN for predicting software effort estimation and benchmark standard data set developed by Lopez et al. are used for training and testing. The consequences are analyzed using MRE, BRE, MMRE, and PRED. It was found that the Generalized Regression NN model delivered better results compared to other ones. The hybrid technique amplifies all the advantages of intelligent gives better results. Tiwari et al. [13] highlighted how to classify mammogram into normal or abnormal. Here, Feed-Forward Artificial Neural Network is used as a classifier for segregating mammogram. Gharehchopogh et al. [1] proposed a hybrid methodology to tune the parameters of COCOMO II, which used COCOMO 81 data set collected from the literature for testing the proposed technique. It was found that the proposed algorithm is able to moderate COOCMO II parameters to estimate an effective effort. Patil et al. [14] proposed a hybrid technique that was the integration of NN and COCOMO II. The Principle Component Analysis is used for mapping exact input values with the proposed technique to predict estimation effort. It was found that outperformed is compared to ANN without PCA. Khatibi et al. [15] proposed a hybrid method which is the integration of fuzzy clustering, analogy, and artificial neural networks to increase the predicted effort estimation of the software product. Hari et al. [16] proposed the CPN hybrid model where C stands for Clustering data using *K*-means, *P* stands for implementing Particle Swarm optimization (PSO) on clusters to secure the COCOMO constant parameter values. Finally, *N* stands for Neural Network trained by backpropagation. The performance of the proposed method is evaluated by MARE and it was outperformed compared to the COCOMO model. Pahariya et al. [17] presented computational intelligence techniques, ensemble techniques, and proposed a new recurrent architecture on genetic programming (RGP) for SCE in this study, tenfold cross validation was performed. All the techniques are tested using ISBSG data set. It was observed that AM ensemble outperformed compared to the others, and proposed methodology outperformed compared to all host techniques. Bedsore et al. [4] proposed an PSO-ABE hybrid method which has outperformed that of the existing methods. Emad et al. [18] investigated and presented a new intelligence paradigm based on the functional network to forecast that is highlighted on numerous software effort estimates elements.

3 Spiking Neural Networks

From experimental studies, it explored that biological neurons use pulses or spikes rather than continuous variables to encode information. More recently, a third-generation [19] neural networks are called Spiking Neural Networks (SNN) have demonstrated power in solving more complex problems, long-term information storage, similar to biological counterparts uses pulses or spikes to represent information flow. In most cases, spiking neuron model implementation based on the basic assumption is spike times which is transmits neural information [5]. In the deterministic model, a spiking neuron p, arises whenever the potential (membrane potential) to reach a certain threshold Θ_p [20]. This potential T_p is the total amount of an Inhibitory Postsynaptic Potential (IPSP) and Excitatory Postsynaptic Potential (EPSP) which is resulting from the other neuron r is firing that are connected through a synapse to neuron p is presented in [21]. At time y, a presynaptic neuron p is firing, generate the potential T_p at the time t is modeled by term $w_{r,p}\ x_{r,p}\ (t\text{-}y)$ where $w_{r,p}$ is weight > 0 and $x_{r,p}\ (t\text{-}y)$ is a response function. Then, a membrane potential is reset. In the literature, the spiking models are Hodgkin–Huxley, Izhikevich, Integrated-and-Fire, and so on. These methods are trying to mimic the above features.

3.1 Leaky Integrated-and-Fire (LIF) Model

From the review papers [22, 23], it was discovered that most widely and commonly used model to represent the spiking neurons based on the principle of electronics and delivered from the Hodgkin-Huxley neuron model is known as Leaky Integrated-and-Fire model [24]. The starting position of the neuron potential is u_{rest}, an excitatory current I (t), and then gradually grows up of the membrane potential voltage $u(t)$ and if the current is not strong enough it leaks the voltage. Whenever the current I(t) is strong enough to grow up the voltage $u(t)$ is exceeded the threshold v, then a spike is generated. This may be represented by the first-order linear differential equation [24].

$$\tau_{\mathrm{m}}\frac{\mathrm{d}u}{\mathrm{d}t} = u(t) - u_{\text{rest}} + RI(t) \tag{1}$$

$$c\frac{\mathrm{d}u}{\mathrm{d}t} = -\frac{1}{R}(u(t) - u_{\text{rest}}) + I(t) \tag{2}$$

where τ_{m} is a membrane constant time, R is the resistance of the membrane. Once it is generating, the spike r which is reset to u_{rest}. The amount of time required to generate a spike is $t^{(f)}$ and the dynamics of the neuron are $t > t^{(f)}$. All these spike signals are binary or none in the neuron as a result of a tremendous analog current flow into every one of it is postsynaptic neurons for $t > t^{(f)}$ which is generally shown as [24]

$$i(t) = w\alpha(t - t^{(f)}) \tag{3}$$

$$\alpha(t) = \left[\exp(-t/\tau_1) - \exp(-t/\tau_2)\right] \tag{4}$$

where two neurons are connecting to each other using synapse of conductance is w or weight and the synaptic constant time is $\tau_1 > \tau_2$. The whole current in a postsynaptic is the positive current flow through the neuron is connected to the weight is $\{w_j\}$ of $n - 1$ presynaptic neurons. Where $j = 0, 2 \ldots n - 1$ is [24]

$$I(t) = \sum_{j=1}^{n} w_j \sum_{f} \alpha\left(t - t_j^{(f)}\right) \tag{5}$$

Here, synaptic transmission delay is $\Delta_{\mathrm{j},} j = 0, 2 \ldots\ n - 1,$ then $t^{(f)}$ is represented by the postsynaptic neuron is $\left(t_j^{(f)} + \Delta_r\right)$. An absolute refractory period Δ_r of the LIF model. Once you reached the threshold in time $t = t_j^f$, the integration takes place again after $\left(t_j^{(f)} + \Delta_r\right)$ from the starting potential u_{rest}.

3.2 Data Encoding

Encoding mechanisms are essential for transforming analog input signals into spikes. There are various encoding techniques for SNN, basically a rate coding or temporal coding. The NueCube have four different encoding mechanisms which are used to represent information into temporal coding [25] the AER, Step-Forward encoding algorithm, moving-window spike encoding algorithm, and Bens Spiker Algorithm. In this paper, the author used to Threshold on ISBSG, CHINA, and IBMDSP generate spike trains athat re entered into the SNN reservoir from the input layers of neuron architecture.

3.3 Learning Mechanism

Spike Time-Dependent Plasticity (STDP) is a representation of competitive Hebbian learning that utilizes relative temporal timing data to adjust the connection weight of a particular neuron between output and input spikes [25]. Experimental results in neuroscience have shown that "synapses grow up their weight if a presynaptic spike reaches just before postsynaptic spike leads to Long-Term Potential (LTP) and decrease if it reaches late leads to Long- Depression Potential (LDP)". Based on this fact, the STDP can lead to stability in LTP and LDP by which postsynaptic neurons sensitive to the timing of spikes and initiates competition among the presynaptic neurons, as a result, increases information propagation through the network, spike synchronization between the neurons and shorter latencies.

4 Research Methodology

In our practice, we are taking the general regression analysis model which is used to predict the relationship between one predicted output variable and one or more predictor input variables The general regression model may be represented as [26]

$$Y = \beta X + \beta_1 + \hat{e} \tag{6}$$

where $X = \{x_1, x_2, x_3 \ldots x_\mathrm{k} \ldots x_\mathrm{n}\}$ is a vector which consists sample of predictor variables, β and β_1 are model coefficients, Y is a predicted output vector, and $\hat{e}$ is an error. The basic idea of regression analysis is to assess the association between dependent and predictor variables. Effort estimation mimics the general regression model that predicts effort estimation. The effort formula can be represented as [27]

$$\text{Effort} = C * \text{Size} + e \tag{7}$$

where C is a Product delivery rate, an effort may be measured by Person-Months and the variable size can be measured in Kilo Line of Code. The project size metric LOC which is unable use in all cases because of it has the shortcomings which are highlighted in [28]. Function Point Analysis (FPA) is an alternative technique to size calculated to overcome the shortcomings of LOC. It breaks the system into smaller parts; they can better analyze and understood. The FPA [28] principle is simply based on the number of "functions" are required to develop software. In this article, we used FPA to calculate size of the given data set.

There are three software effort estimation datasets obtained (1) China dataset from public domain [29], (2) IBMDSP dataset from the IBM Data Processing Services (DPS) organization [30], and (3) ISBSG dataset from the ISBSG repository [27]. The min-max normalization technique is applied to normalized the datasets. Once the data is normalized, which can be made into samples where numbers of tuples considered as an equal sample. Each tuple included feature and the target value is corresponding to the sample where the mean value of all the actual effort values of their sample. The proposed technique has implemented using the most popular prediction tool is known as NeuCube tool used to predict the cost of any problem, which is depicted in Fig. 1. It has different components which are given in [6]. The first component of

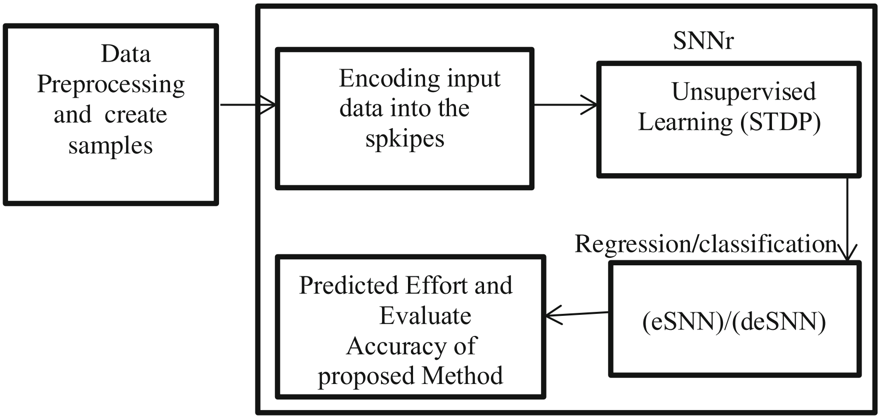

Data Preprocessing and create samples
SNNr
Encoding input data into the spkipes
Unsupervised Learning (STDP)
Regression/classification
Predicted Effort and Evaluate Accuracy of proposed Method
(eSNN)/(deSNN)

Fig. 1 Proposed methodology to enhance accuracy of software cost estimation

the proposed methodology is encoded, which takes an input data samples converted into positive and negative binary (trains of spikes) using an encoding technique. The output has fed into the training of 3D SNN reservoir (SNNr) then output of the SNNr is sent to the classifier/regression (e.g., eSNN, deSNN) which maps SNNr states into predictive output values. In the Neucube, learning process is two phase. In the first phase, unsupervised way implementation using Spike Time-Dependent Plasticity rule (STDP) [31], where temporal data are entered into the relevant area of SSN classifier over time.

Before performing unsupervised learning to set the learning parameters are potential leak rate, threshold of firing, refractory time, STDP rate, and LDC probability to modify the initial set connection weights. In the case of spiking neural network, the weights are adjusted iteratively to minimize the error in the modeling process. SSN reservior (SNNr) learn from input data which is same data in groups of neurons. The second is a supervised learning which takes the output of trained spikes as input, perform regression to predict predefine output spike sequences. The next step is to verify the accuracy of the model built by deSNN learning. In this paper, performance of the proposed technique evaluate using statistical measures.

Mean Magnitude of Relative Error (MMRE) for each observation i can be obtained as [32]

$$\text{MMRE} = \sum_{i}^{n} \frac{\hat{E}_i - E_i}{E_i} \tag{8}$$

where E_i is Actual effort, $\hat{E}_i$ is Estimated Effort.

Root Mean Square Error (RMSE) for each observation i can be obtained as [33]

Table 1 A statistical description of three datasets

Dataset	Features	Observations	Min effort	Max effort	Mean effort	Standard deviation effort
IBMDSP	07	24	0.5	105.2	21.87	28.41
ISBSG-10	105	4106	4	645,694	5356.38	19,789.68
CHINA	19	499	26	54,620	3921.04	6480.85

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(E_i - \hat{E}_i\right)^2} \tag{9}$$

In this research article, the above-proposed methodology is applied to software Engineering domain where to predict the effort estimation of the software at the earliest stage of the lifecycle of a process. Here, estimation data samples are encoded using Threshold technique which has fed into the training process of the SNN. After training to predict effort estimation by giving the test sample. Once known, the effort it easily calculates the cost, time, and budget required to complete the project. Finally, calculate the accuracy of the proposed system should be minimized which is described in session 5. The project manager is using the above process to estimate cost, time, effort and budget for upcoming projects effectively.

5 Results and Discussions

We conducted experiments with the Nuecube tool [34] is an open-source implementing SNN for simulating experiments on different datasets, viz. ISBSG, IBM, and China are presented in the Table 1. All the data sets are divided into samples with equal number of tuples and at the same time, the target value is mean of all actual effort corresponding to the samples. While conducting experiments the suitable parameters such as Potential Leak Rate, k, and Spike Threshold are considered for tuning parameters of effort estimation to get an accurate estimation of the software cost.

After conducting experiments (see Table 2), it was examined that in the case of ISBSG data set, the RSME is 0.01 for all three cases of $k = 6$, 4, and 2, and Potential Leak Rate is 0.002, Spike Threshold is 0.05 constant, and MMRE values are 0.03, 0.50, and 1.61 (Tables 3 and 4). Further, we changed the Potential Leak Rate to 0.001 (see Table 5), the rest of the parameters are same for all three cases, the RMSE should be same and MMRE are 0.18, 0.51, and 1.61, respectively.

Subsequently, (see Table 3) in the case of CHINA data set, the RSME is 0.02 for all three cases of $k =$ 6, 4, and 2, and Potential Leak Rate is 0.002, Spike Threshold is 0.05 are constant, MMRE values are 0.16, 0.51, and 0.23. Further, we changed the Potential Leak Rate to 0.001(see Table 6), the rest of the parameters are same for all

Table 2 RMSE, MMRE values, and K parameters for ISBSG dataset [27]

Employed method	*K*	Potential leak rate	Spike threshold	RMSE	MMRE
SNN	6	0.002	0.5	0.01	0.03
	4	0.002	0.5	0.01	0.50
	2	0.002	0.5	0.01	1.61

Table 3 RMSE, MMRE values, and *K* parameters for CHINA dataset [29]

Employed method	*K*	Potential leak rate	Spike threshold	RMSE	MMRE
SNN	6	0.002	0.5	0.02	0.16
	4	0.002	0.5	0.02	0.05
	2	0.002	0.5	0.02	0.23

Table 4 RMSE, MMRE values, and K parameters for IBMDSP dataset [35]

Employed method	*K*	Potential leak rate	Spike threshold	RMSE	MMRE
SNN	6	0.002	0.5	0.02	0.17
	4	0.002	0.5	0.02	0.07
	2	0.002	0.5	0.02	0.25

Table 5 After modifying Potential Leak Rate, RMSE, MMRE values, and *K* parameters for ISBSG dataset [27]

Employed method	*K*	Potential leak rate	Spike threshold	RMSE	MMRE
SNN	6	0.001	0.5	0.01	0.18
	4	0.001	0.5	0.01	0.51
	2	0.001	0.5	0.01	1.61

three cases, the RMSE should be same $k = 6$, 2 but $k = 4$ the RSME is 0.03, and MMRE are 0.16, 0.05, and 0.23, respectively.

Similarly, (see Table 4) in the case of the IBM data set, the RSME is 0.02 for all three cases $k = 6$, 4, and 2, and Potential Leak Rate is 0.002, Spike Thershold is 0.05, and MMRE values are 0.17, 0.07, and 0.25. Further, we changed the Potential

Table 6 After modifying Potential Leak Rate, RMSE, MMRE values, and *K* parameters for CHINA dataset [29]

Employed method	*K*	Potential leak rate	Spike threshold	RMSE	MMRE
SNN	6	0.001	0.5	0.02	0.16
	4	0.001	0.5	0.02	0.05
	2	0.001	0.5	0.03	0.23

Table 7 After modifying Potential Leak Rate, RMSE, MMRE values, and *K* parameters for IBMDSP dataset [30]

Employed method	*K*	Potential leak rate	Spike threshold	RMSE	MMRE
SNN	6	0.001	0.5	0.01	0.18
	4	0.001	0.5	0.01	0.08
	2	0.001	0.5	0.01	0.25

Table 8 Comparison of RMSE [27] on ISBSG

S. No.	Employed method	RMSE value
1	**SNN**	**0.01000**
2	RGP	0.03275
3	GP-RGP	0.03345
4	GP-GP	0.04676
5	GMDH-GP	0.03098
6	GP-GMDH	0.04833

Table 9 Comparison of MMRE [35] on IBM

S. No.	Employed method	MMRE value
1	**SNN**	**0.15000**
2	PSO-ABE	0.26000

Leak Rate to 0.001(see Table 7), the rest of the parameters are same for all three cases, the RMSE should be is 0.01 and MMRE are 0.18, 0.08, and 0.25 respectively.

Further, from the experiments (see Tables 8 and 9); we found that the present outcomes are much better compared to the other ones.

6 Conclusion

In this article, the proposed technique is used to predict the accurate process estimated cost of the software at the primary stages of the software process lifecycle and to evaluate the performance of proposed technique based on three estimation benchmark datasets are IBM, ISGSB, and CHINA using statistical measures are RMSE and MMRE. Further, it compared the performance of the SSN with hybrid techniques that are given in Tables 8 and 9. We found that the performance of the SNN in terms of RMSE and MMRE values was better than other hybrid models. Hence, we conclude that after an extensive investigation is that the SNN model is comparatively best predicted among all the other techniques.

References

1. Gharehchopogh, F.S., Pourali, A.: A new approach based on continuous genetic algorithm in software cost estimation. J. Sci. Res. Dev. **2**(4), 87–94 (2015)
2. Haykin S.: Neural networks—a comprehensive foundation, 2nd ed. In: Upper Saddle River, Prentice-Hall, NJ, USA, pp. 156–255 (1999)
3. Gerstner, W.: Time structure of the activity in neural network models. Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top **51**(1), 738–758 (1995)
4. Maass, W., Zador, A.: Computing and learning with dynamic Synapses. Pulsed Neural Netw. **6**, 321–336 (1999)
5. Gerstner, W., Kistler, W.M.: Spiking Neuron Models: Single Neurons, Populations, Plasticity. Cambridge Univ. Press, Cambridge, MA, USA (2002)
6. Thorpe, S., Fize, D., Marlot, C.: Speed of processing in the human visual system. Nature **381**(6582), 520–522 (1996)
7. Schliebs, S., Kasabov, N.: Evolving spiking neural network—a survey. Evolving Syst. **4**(2), 87–98 (2013)
8. Kasabov, N., Dhoble, K., Nuntalid, N., Indiveri, G.: Dynamic evolving spiking neural networks for on-line spatio- and spectro-temporal pattern recognition. Neural Netw. **41**, 188–201 (2013)
9. Mohemmed, A., Schliebs, S., Matsuda, S., Kasabov, N.: Training spiking neural networks to associate spatio-temporal input–output spike patterns. Neurocomputing **107**, 3–10 (2013). https://doi.org/10.1016/j.neucom.2012.08.034
10. Kasabov, N.: NeuCube: a spiking neural network architecture for mapping, learning and understanding of spatio-temporal brain data. Neural Netw. **52**, 62–76 (2014)
11. Attarzadeh, I., Mehranzadeh, A., Barati, A.: Proposing an enhanced artificial neural network prediction model to improve the accuracy in software effort estimation. In: IEEE Fourth International Conference on Computational Intelligence, Communication Systems and Networks (CICSyN), pp. 167–172 (2012)
12. Ghose, M.K., Bhatnagar, R., Bhattacharjee, V.: Comparing some neural network models for software development effort prediction. In: IEEE 2nd National Conference on In Emerging Trends and Applications in Computer Science (NCETACS), pp. 1–4 (2011)
13. Tiwari A., Bhateja V., Gautam A., Satapathy S.C.: ANN-based classification of mammograms using nonlinear preprocessing. In: 2nd International Conference on Micro-Electronics, Electromagnetics and Telecommunications. Lecture Notes in Electrical Engineering, vol. 434. Springer, Singapore (2017)
14. Patil, L.V., Waghmode, R.M., Joshi, S.D., Khanna, V.: Generic model of software cost estimation: a hybrid approach. In: IEEE International Advance Computing Conference (IACC), pp. 1379–1384 (2014)
15. Bardsiri, V.K., Jawawi, D.N.A., Hashim, S.Z.M., Khatibi, E.: Increasing the accuracy of software development effort estimation using projects clustering. IET Software **6**(6), 461–473 (2012)
16. Hari, C.V., Sethi, T.S., Kaushal, B.S.S., Sharma, A.: CPN-a hybrid model for software cost estimation. In: IEEE on Recent Advances in Intelligent Computational Systems (RAICS), pp. 902–906, (2011)
17. Pahariya, J.S., Ravi, V., Carr, M.: Software cost estimation using computational intelligence techniques. World Congress on Nature & Biologically Inspired Computing, pp. 849–854 (2009)
18. Paugam-Moisy, H., Bohte, S.: Computing with spiking neuron networks. In: Handbook of natural computing, Springer, pp. 335-376. Berlin Heidelberg (2012)
19. Gupta, A., Long, L.N.: Character recognition using spiking neural networks. IEEE Int. Joint Conference on Neural Networks, IJCNN **2007**, 53–58 (2007)
20. Maass, W.: Networks of spiking neurons: the third generation of neural network models. Neural Networks **10**(9), 1659–1671 (1997)
21. Reid, D., Hussain, A.J., Tawfik, H.: Financial time series prediction using spiking neural networks. PloS one **9**(8) (2014)

22. Prasad, C., Saboo, K., Rajendran, B.: Composer classification based on temporal coding in adaptive spiking neural networks. In: IEEE International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2015)
23. Zhang, Z., Wu, Q., Wang, X., Sun, Q.: Training spiking neural networks with the improved Grey-Level Co-occurrence Matrix algorithm for texture analysis. In: 11th International Conference on Natural Computation (ICNC), pp. 1069–1074 (2015)
24. Paugam-Moisy, H., Bohte, S.: Computing with spiking neuron networks, pp. 335–376. In Handbook of natural computing, Springer, Berlin Heidelberg (2012)
25. Bose, P., Kasabov, N.K., Bruzzone, L., Hartono, R.N.: Spiking neural networks for crop yield estimation based on spatiotemporal analysis of image time series. IEEE Trans. Geosci. Remote Sens. **54**(11), 6563–6573 (2016)
26. Campbell, D., Campbell, S.: Introduction to regression and data analysis. In: Statlab Workshop, (2008)
27. ISBSG: The international software benchmarking standards group, http://www.isbsg.org (2011)
28. Kaur, M., Sehra, S. K.: Particle swarm optimization based effort estimation using Function Point analysis. In: IEEE 2014 International Conference on In Issues and Challenges in Intelligent Computing Techniques (ICICT), pp. 140–145 (2014)
29. http://promise.site.uottawa.ca/SERepository
30. Matson, J.E., Barrett, B.E.: Software development cost estimation using function points. IEEE Trans. Software Eng. **20**(4), 275–287 (1994)
31. Hussain, A.J., Reid, D., Tawfik, H.: A spiking neural network for financial prediction
32. Kolodner, J.: Case-based reasoning. Morgan Kaufmann (2014)
33. Huang, X., Capretz, L.F., Ren, J., Ho, D.: A neuro-fuzzy model for software cost estimation. In: IEEE Third International Conference on Quality Software, 2003. Proceedings, pp. 126–133 (2003)
34. https://kedri.aut.ac.nz/KEDRI-R-and-D-Systems/neucube
35. Bardsiri, V.K., Jawawi, D.N.A., Hashim, S.Z.M., Khatibi, E.: A PSO-based model to increase the accuracy of software development effort estimation. Software Qual. J. **21**(3), 501–526 (2013)

Modeling of PV Powered Seven-Level Inverter for Power Quality Improvement

R. Arulmurugan and A. Chandramouli

Abstract In this article, a solitary-phase seven-level series connected H-Bridge powered by photovoltaic MPPT-based SHAPF in view of basic controller is proposed. SRF is utilized for reference input current extraction and to create pulses for the SHAPF. The principle point of the cascaded bridge is to dispense harmonics, enhance power factor, and reactive energy compensation of the single-phase distribution framework. The suggested control calculation has two parts, changing the load current into stationary reference outline directions and estimation of peak amplitude of load currents. Consequently, a basic and dependable controller effortlessly of execution was created. The calculation for single-phase SHAF is intending to perform with exact tracking performance under step changes in load currents and to give great dynamic compensation. In this article, synchronous reference theory PLL with Inverse-Park change is adopted for producing quadrature part of current. The execution of the control calculation is tried and assessed utilizing MATLAB/Simulink tool.

1 Introduction

In last years, harmonics is the most essential issue as far as power quality because of widespread of energy conversion or power electronic gadgets in business, mechanical, and residential loads. In dissemination frameworks, the usage of nonlinear loads, for example, PCs, variable/flexible speed drives, light-emitting device frameworks, and conservative fluorescent lamps, and so onward is utilizing generally and inclined to harmonics [1]. These harmonics are causing serious issues, for example, control power losses in equipment's, breaking down or malfunctioning of gadgets, harming of delicate loads, and drive engine disappointments. Subsequently, it is a genuine

R. Arulmurugan · A. Chandramouli (✉)
Department of EEE, S R Engineering College, Warangal, Telangana, India
e-mail: chandersrec@gmail.com

R. Arulmurugan
e-mail: arul.lect@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_12

concern in distribution frameworks for both purchasers and providers to dispose of harmonics and meet the necessities of IEC 61000-3-2 or IEEE 519-1992 [2].

The harmonics created by the loads are making grid voltages be mutilated. Traditionally, detached filters are utilized for harmonic alleviation and reactive power compensation. In any case, these experience the ill effects of weaknesses like massiveness, cost, resonance and fixed remuneration [3]. In such a way, a dynamic arrangement is favored that fits the compensation is a shunt active filter. The role of APF is to compensate reactive power and harmonic currents with enhanced power factor delivered by the load. The controller needs to path the progression changes in the load precisely and to choose reference current appropriately for better compensation.

Keeping reliability and accuracy in view, numerous methods are explained in surveys for quadrature signal production. ZCD technique [4] is straightforward at the same time, sensitive to variations of grid. In general the utilized technique is SRFR and SOGI-based hypothesis [5, 6]. It is less precise to unequal and brings down harmonic components. Be that as it may, SRF hypothesis with reverse stop change based calculation is discovered palatable under contorted conditions with low computational burden. In any case, application and usage of this control process for a five-level cascaded H-connect dynamic power filters has not increased much consideration in the literature.

MLI has increased much consideration in virtue of its gigantic preferences over traditional voltage source inverters. The traditional two-level inverter is likewise fit for taking care of harmonic reduction, power factor change, and reactive power under different load changes. Yet, because of progression of electronic gadgets and controllers, MLIs have demonstrated their capacity to compensate issues of power quality with straightforwardness, ease, dependability, and high-quality output. There are numerous methods suggested in the literature [7, 8]. Flying capacitor-based inverter, neutral point clamped inverter, and cascaded H-connected type MLI are discovered reasonable for SHAPF application effortlessly of control.

In this article, cascade H-bridge sort of MLI is utilized to lessen the ranking of the gadgets utilized and disposal of harmonics with an expansion in levels of the converter. This technique likewise lessens the exchanging misfortunes and declines the ratings of the direct current interface capacitors utilized. The control calculation is discovered effective in linear/nonlinear and increment in load circumstances. With a specific end goal to every one of these topologies, different PWM strategies were likewise suggested in the surveys which incorporates specific harmonic disposal based PWM, Carrier-based PWM, Multilevel space vector based PWM and so on [9, 10]. The fundamental favorable position of this CHB inverter is expanding of switching levels by increasing the number of H-connected in the circuit. This paper utilizes a basic SRF-based control in with reverse park alteration to generate quadrature signal for reactive power compensation and harmonic minimization.

A cascade H-bridge-based SHAF is proposed in this article, nonlinear load cases under steady state and dynamic conditions are completed utilizing Simulink, simpower frameworks block set and its execution discovered palatable.

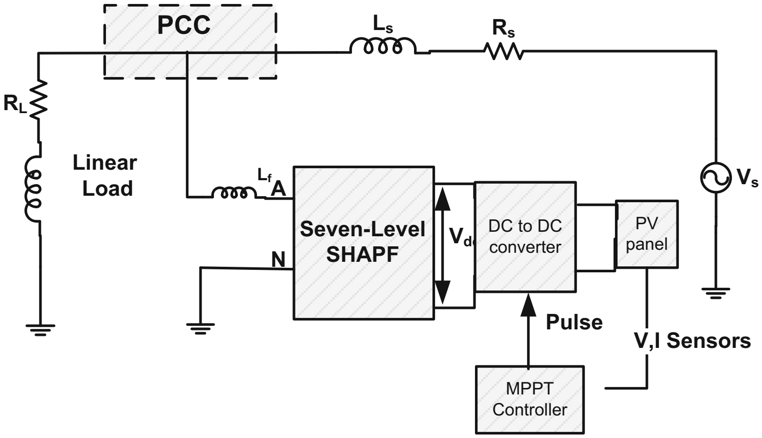


Fig. 1 Line chart of the proposed system

2 Designed Configuration and Controller

The Cascade H converter-based SHAPF appeared in Fig. 1. Every H-connect comprises of a two-leg VSC comprising of four IGBT switches. There are two H-connect VSCs are utilized for producing seven-level yield over the inverter was used. SHAF is associated in middle of source and load in parallel through an interfacing inductor L_f at the PCC. The recommended controller for SHAF is equipped of keeping up the THD within the limits by removing the harmonics in the input or grid current. Reactive power, Power factor rectification, and harmonic compensation are likewise done even under changing nonlinear and linear load situations to examine the execution of the controller. The SHAF can be worked with required dynamic and responsive power infusion by modifying the greatness and phase of the system. The ratings of the designed framework are listed in reference section. The rating of the SHAF ought to be 15% more than the Load rating for more secure and monetary operation.

2.1 Srf-Pll

At present, the essential PLL topology generally utilized in all the area of research is SRF-based PLL. V represents as input voltage signal and is considered as $V\alpha$ and its orthogonal part that is moved by 90° is $V\beta$. These two parts ($V\alpha$ and $V\beta$) are in constant reference frame and these are changed over to synchronous turning reference frame (d, q) by utilizing Park's transformation as represent in Eq. (1) [11].

$$T = \begin{bmatrix} \cos\hat{\theta} & \sin\hat{\theta} \\ -\sin\hat{\theta} & \cos\hat{\theta} \end{bmatrix} \tag{1}$$

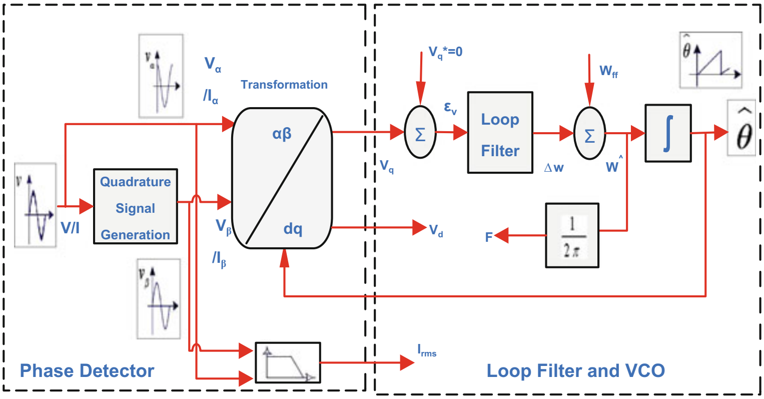


Fig. 2 Control diagram of SRF-PLL

The d-q parts in its reference outline are controlled by a precise position with a feedback signal related with it. The grid voltage amplitude is thoroughly relating with the d-segment and corresponds with all its d-segment and furthermore making q-part to zero. The transformation yield is gone through a circle channel (LF) for taking out any high-frequency noises within it and afterward included with the nominal frequency and afterward provided to a VCO to create central stage point θ. So, as to get a correct amplitudes with an adjusted arrangement of quadrature yields and in-phase, the frequency produced by the phase-locked loop ought to be equivalent to the input fundamental frequency ($W_{\text{ff}} = 2\pi * 50$). Proportional-Integral controller is the essential loop filter utilized as a part of all these PLL topologies.

2.2 Inverse-Park Transformation

Figure 2 demonstrates the structure of inverse-park transformation. Park transformation is done (i.e., $\alpha\beta 0/dq0$ s) and these yields are utilized for opposite park alteration as appeared in Fig. 2. The elements of the phase indicator predominantly relies the low-pass filter that is utilized after the transformation to filter out any noises or harmonics that are available in V_d and V_q.

2.3 Reference Current Generation

The peak amplitude of active segment of current is computed as appeared in Fig. 3. The load current is detected and provided to inverse-park transformation to create quadrature signals ($I_L\alpha$ and $I_L\beta$) and afterward changed back to I_d and provided to a low-pass filter. The yield is then included with the output created by the DC voltage

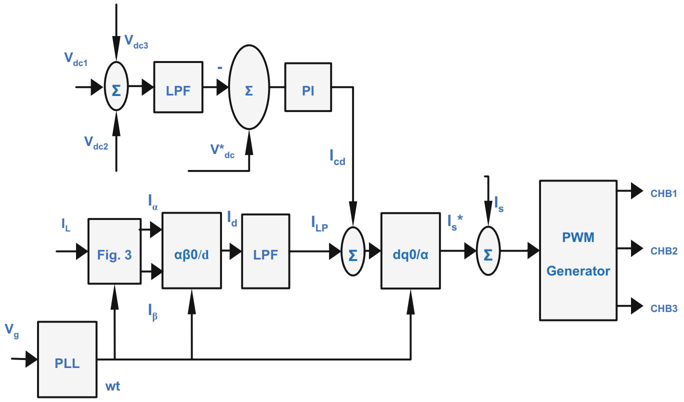


Fig. 3 Control modes of RSC Generation

control circle to deliver reference dynamic part of current ($I_{LP} + I_{CD}$). The deliberate voltage (V_{dc}) over the two DC capacitors are summed and contrasted and the direct current bus reference voltage (V_{dc}^{*}). The error of the signal at *n*th examining moment is given in Eq. (2):

$$V_{\mathrm{d}}(n) = V_{dc}^{*}(n) - V_{dc}(n) \tag{2}$$

The voltage error V_{d}(n) is then provided to Proportional-Integral controller to direct the DC bus voltage of SHAPF. At *n*th examining moment, the yield of the PI controller is represented in Eq. (3) as

$$I_{cd}(n) = I_{cd}(n-1) + k_p\{V_{\mathrm{dcer}}(n) - V_{\mathrm{dcer}}(n-1)\} + k_i V_{\mathrm{dcer}}(n) \tag{3}$$

where K_p and K_i are proportional gain and integral gains of the PI controller. $V_{\mathrm{dcer}}(n)$ and $V_{\mathrm{dcer}}(n\text{-}1)$ are the direct current bus voltage errors in *n*th and (*n*-1)th moment and $I_{cd}(n)$ and $I_{\mathrm{cd}}(n\text{-}1)$ are the amplitudes of dynamic segment of currents at the basic reference current in *n*th and (*n*-1)th moment.

The magnitude current (I_{LP}) and the yield of the proportional-integral controller (I_{cd}) are summed up to altered *Iα* (source current reference) from *dq*0 part and afterward contrasted with the genuine source or grid current to create error magnitude of current and after that provided to a pulse width modulation controller to produce pulses to MLI.

The photovoltaic are demonstrated as nonlinear sources voltage. The sources are associated with direct current to direct current converters which are joined at the direct current side of a inverter (DC/AC). The DC/DC associated with the PV array operates as a MPPT controller. Numerous MPPT calculations have been offered in the script, for example, Incremental Conductance (INC), Constant Voltage (CV), Perturbation and Observation (P&O). The P&O technique has been broadly utilized in view of its basic and simple feedback structure and less measured parameters

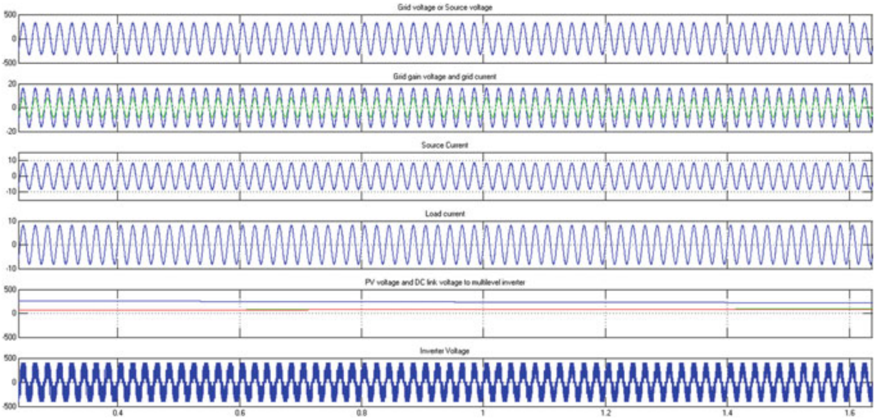


Fig. 4 Performance of seven-level cascaded H-connected inverter with linear load

[12]. The P&O calculation with control input control [13, 14]. As PV voltage and current are resolved, the power is computed. At the MPP, the derivative (dP/dV) is equivalent to zero. The greatest powerpoint can be accomplished by changing the reference voltage by the measure of ΔVref.

3 Simulation Test and Results

In this segment, the designed control calculation is assessed and tried utilizing MATLAB/Simulink on a solitary-phase distribution framework loaded with nonlinear and linear loads. Settled time step of 20 μs with ode3 (Bogacki—Shampine) solver is decided for recreation [15]. Scarcely any experiments are performed for assessments of SHAPF are: The execution of the controller when a nonlinear and linear load is connected is appeared in Figs. 4 and 5 when time period t = 0.4–0.5 s. All these experiments are executed under sinusoidal grid conditions [16, 17]. The execution under powerful conditions is discovered attractive [18]. The direct current bus voltage control, harmonic compensation and reactive power with Power factor change demonstrates the viability of the controller [19]. THD (%) of cases specified above are exhibited in Table 1 demonstrate the viability of the controller [20, 21]. Figure 6 proves the seven-level output of multilevel inverter. Followed by, Fig. 6 shows the pulse chart of single bridge system. Each bridge system used the four switches, THD level of nonlinear waveforms was displaced in Fig. 7. The nonlinear and linear of seven-level THD values were associated with PV seven-level inverter as exposed in Table 1. Table 1 proves that the seven-level inverter reduced the harmonics level more than 15% than seven-level system.

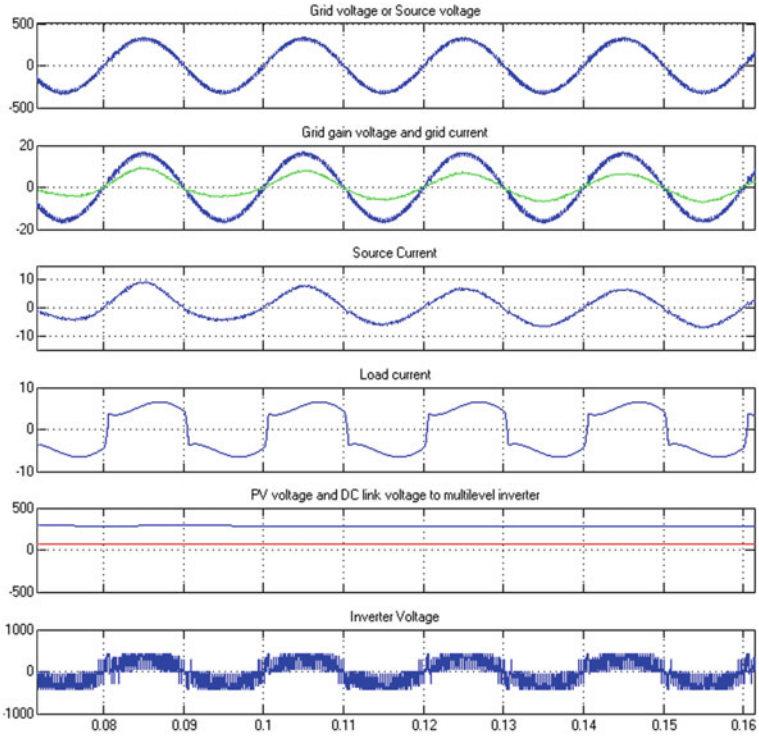


Fig. 5 Performance of seven-level cascaded H-connected inverter with non-linear load

Table 1 THD (%) of test cases

No. of switching converter	THD level	
Type of load	Linear load	Nonlinear load
Seven-level converter	0.63%	34.24%
Seven-level PV system	0.70%	70.25%
Fundamental (50 Hz)=	325.4	143.9

4 Conclusion

In this article, a basic and compelling control calculation in view of Synchronous Reference Frame (SRF) hypothesis for single-phase system with photovoltaic MPPT powered cascaded seven-level H-connected dc to ac converter has been analyzed, exhibited utilizing MATLAB/Simulink tool. This hypothesis is embraced to work in sinusoidal grid voltage conditions and nonlinear load situations. The source or grid current harmonics THD (%) is kept up IEEE 519-1992 breaking points. The control calculation is tremendously encouraging and simple to execute due to its straightforward structure and exactness. The reactive current, harmonic compensation and

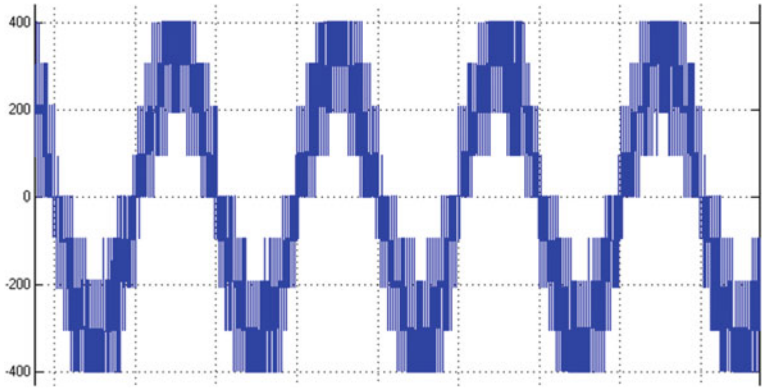


Fig. 6 Seven-level cascaded H-connected bridge output voltage waveform

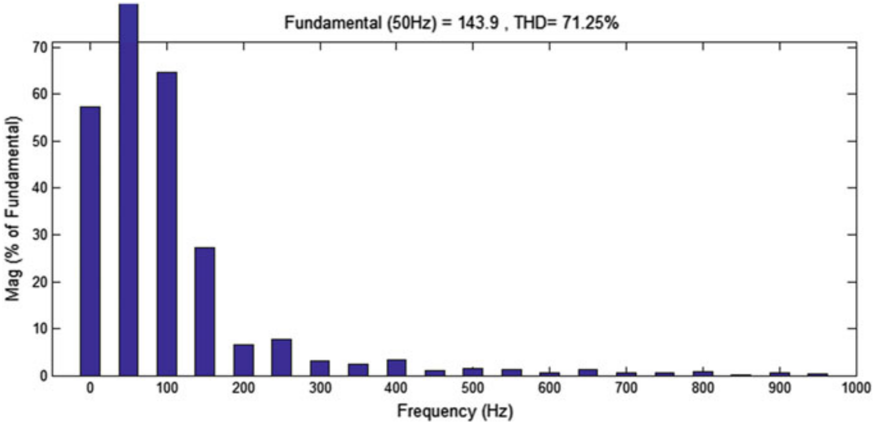


Fig. 7 THD examination of the designed PV seven-level inverter

power factor are effectively done under all relentless-state and dynamic situations. The infused current of the shunt active device of power factor was likewise near the reference esteems and demonstrated a smooth and dependable profile. This paper is further expanded to change the multi fuzzy-based MPPT system.

References

1. IEEE recommended practices and requirements for harmonic control in electrical power systems, IEEE Std. **519** (1992)
2. Electromagnetic compatibility (EMC)—Part 3–2: limits—limits for harmonic current emissions (Equipment Input Current = 16 A per Phase), IEC 61000-3-2 (2005)

3. Timbus, A., Liserre, M., Teodorescu, R., Blaabjerg, F.: Synchronization methods for three phase distributed power generation systems: an overview and evaluation. PESC 2005, pp. 2474–2481 (2005)
4. Kaura, Blasco, V.: Operation of a phase locked loop system under distorted utility conditions. IEEE Trans. Ind. Appl. **33**(1), 58–63 (1997)
5. Hareesh Kumar, Y., Murthy, M.S.R.: A new topology and control strategy for extraction of reference current using single phase SOGI-PLL for three-phase four-wire Shunt Active Power Filter. In: 2014 IEEE International Conference on Power Electronics Drives and Energy Systems (PEDES) (2014)
6. Arulmurugan, R., Venkatesan, T.: Research and experimental implementation of a CV-FOINC algorithm using MPPT for PV power system. J. Electr. Eng. Technol. **10**(1), 30–40 (2015)
7. Law, K.H., Dahidah, M.S.A., Marium, N.: Cascaded multilevel inverter based STATCOM with power factor correction feature. In: Proceedings of IEEE Conference Sustainable Utilization Developing Engineering Technology, pp. 1–7 (2011)
8. Peng, F.Z., Lai, J.S.: Dynamic performance and control of a static var generator using cascade multilevel inverters. IEEE Trans. Ind. Appl. **33**(3), 748–755 (1997)
9. Cheng, Y., Chang, Q., Mariesa, L.C., Pekarek, S., Aticitty, S.: A comparison of diode-clamped and cascaded multilevel converters for a STATCOM with energy storage. IEEE Trans. Ind. Electron. **53**(5), 1512–1521 (2006)
10. Naderi, R., Rahmati, A.: Phase-shifted carrier PWM technique for general cascaded inverters. IEEE Trans. Power Electron. **23**(3), 1257–1269 (2008)
11. Yao, W., Hu, H., Lu, Z.: Comparisons of space-vector modulation and carrier-based modulation of multilevel inverter. IEEE Trans. Power Electron. **23**(1), 45–51 (2008)
12. Rodriguez, J., Moran, L., Correa, P., Silva, C.: A vector control technique for medium-voltage multilevel inverters. IEEE Trans. Ind. Electron. **49**(4), 882–888 (2002)
13. Arulmurugan, R., et al.: Tracking of photovoltaic power system with new Fuzzy Logic Control strategy. J. Electr. Eng. (JEE) vol. 14, edn. 4, pp. 1–10. ISSN 1582-4594 (2014)
14. Gupta, K., Khambadkone, A.M.: A general space vector PWM algorithm for multilevel inverters, including operation in over modulation range. IEEE Trans. Power Electron. **22**(2), 517–526 (2007)
15. Arulmurugan, R., et al.: Improved fractional order VSS inc-cond MPPT algorithm for Photovoltaic Scheme. Int. J. Photoenergy, vol. 2014, Article ID. 128327, 10 pages, (2014)
16. Dahidah, M.S.A., Agelidis, V.G.: Selective harmonic elimination PWM control for cascaded multilevel voltage source converters: A generalized formula. IEEE Trans. Power Electron. **23**(4), 1620–1630 (2008)
17. Ozpineci, B., Tolbert, L.M., Chiasson, J.N.: Harmonic optimization of multilevel converters using genetic algorithms. In: Proceedings Power Electronics, Specialists Conference, pp. 3911–3916 (2004)
18. Agelidis, G., Balouktsis, A., Dahidah, M.S.A.: A five-level symmetrically defined selective harmonic elimination PWM strategy: analysis and experimental validation". IEEE Trans. Power Electron. **23**(1), 19–26 (2008)
19. Arulmurugan, R., et al.: Intelligent fuzzy MPPT controller using analysis of DC-DC buck converter for PV power system applications. In: IEEE International conference on PRIME 2013, on Feb 22–23 (2013)
20. Dahidah, M.S.A., Konstantinou, G., Agelidis, V.G.: SHE-PWM and optimized DC voltage levels for cascaded multilevel inverters control. In: Proceedings IEEE Symposium on Industrial Electronics Applications, pp. 143–148 (2010)
21. Law, K.H., Dahidah, M.S.A., Agelidis, V.G.: SHE-PWM cascaded multilevel converter with adjustable DC sources control for STATCOM applications. In: Proceedings IEEE 7th International Power Electronics. Motion Control Conference, pp. 330–334 (2012)

Social Group Optimization and Shannon's Function-Based RGB Image Multi-level Thresholding

R. Monisha, R. Mrinalini, M. Nithila Britto, R. Ramakrishnan and V. Rajinikanth

Abstract In recent years, gray and RGB image multi-level preprocessing practices are extensively discussed by the researchers because of its practical importance. In this work, an approach based on the integration of Shannon's function and Social Group Optimization (SGO) are proposed to find optimal thresholds for a class of benchmark RGB pictures. A novel Cost Function (CF) based on the entropy value, PSNR, and SSIM are proposed to guide the SGO-assisted threshold search procedure. The experimental work is implemented using Matlab software and the performance of this practice is evaluated by computing the image quality measures. Finally, the superiority of the proposed approach is validated using PSO, BFO, FA, and BA algorithm-based results. This study confirms that the average PSNR and SSIM obtained with the SGO and Shannon are better compared with the other algorithms adopted in this paper.

1 Introduction

Image processing approaches are widely adopted in a variety of engineering and medical domains to process the picture frames in order to extract the vital information. Image multi-level thresholding is one of the image preprocessing procedures widely implemented by the researchers to examine the grayscale [1–3] and RGB pictures [4, 5]. This approach will enhance/extract the information of the picture by clustering the similar pixel levels based on the chosen threshold values.

Conventional thresholding procedures are widely adopted to process the picture frame based on the bi-level or tri-level thresholding approach. When the chosen threshold increases, the conventional approach requires more time to identify the

R. Monisha · R. Mrinalini · M. Nithila Britto · V. Rajinikanth (✉)
Department of Electronics and Instrumentation, St. Joseph's College of Engineering, Chennai 600 119, Tamilnadu, India
e-mail: rajinikanthv@stjosephs.ac.in

R. Ramakrishnan
PRP Division, Bhabha Atomic Research Centre, Kalpakkam 603 102, Tamilnadu, India

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_13

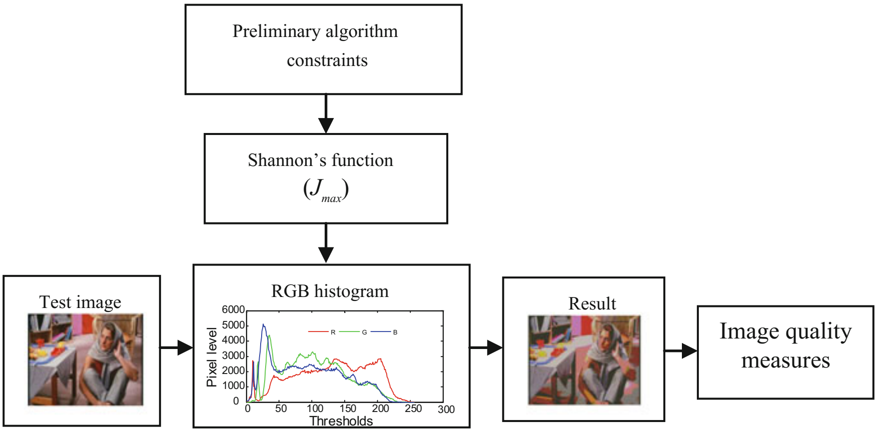


Fig. 1 Implementation of the proposed approach

best possible threshold values. Hence, soft computing algorithm guided techniques are implemented to preprocess the image frames [6, 7]. Soft computing approaches are also efficient in offering the better threshold levels with increase image quality measures [8–10]. Soft computed integrated approaches like the maximization of between-class variance and maximization of the entropy value are already implemented to enhance the information of the grayscale and RGB pictures [11, 12]. In this paper, Shannon's entropy based technique is considered to find the optimal threshold for a class of benchmark RGB images widely discussed in the literature [4, 5]. In order to reduce the computational complexity, recently developed heuristic approach known as the Social Group Optimization (SGO) is considered to guide the threshold search. This paper also proposed a novel Cost Function (CF) for the optimization search by integrating the maximal Shannon's entropy, PSNR, and SSIM. The performance of the proposed approach is then validated against the PSO-, BFO-, FA-, and BA-assisted thresholding techniques available in the literature [13, 14]. Experimental results of this study confirm that proposed SGO + Shannon's offers better PSNR and SSIM values compared with other approaches considered in this paper.

2 Methodology

This section present s the information of techniques considered in this work to preprocess RGB image frames. Figure 1 depicts the procedure implemented in this paper.

Figure 1 depicts the pictorial illustration of the procedure considered to threshold the test image using the considered heuristic algorithm. This technique searches the optimal threshold in the RGB histogram till $J_{\max}$ is reached. Finally, the image quality measures for the result are recorded to evaluate the performance of the heuristic algorithms considered in this work.

2.1 Shannon's Entropy

In image processing applications, entropy assisted procedures are widely adopted to enhance the irregularity in the picture frame. In this paper, Shannon's entropy-based procedure is implemented to preprocess the RGB image its information can be found in the following work [15, 16]. In this paper, optimal thresholds for R, G, and B levels are obtained separately.

The Shannon's function for an individual image frame can be expressed as follows:

Let $M*N$ represent the dimension of an image frame, then, the pixel with coordinates (x, y) for R/G/B level can be represented as $f\ (x, y)$, for $x \in \{1, 2, \ldots, M\}$ and $y \in \{1, 2, \ldots, N\}$.

Let L be the number of thresholds of picture frame $'I_0'$ and the set of all threshold values$\{0, 1, 2, \ldots, L-1\}$ can be indicated as G, in such a way that

$$f(x, y) \in G\, \forall (x, y) \in \text{image} \tag{1}$$

If $H = \{h_0, h_1, \ldots, h_{L-1}\}$ denotes the normalized histogram of R/G/B level, then for a chosen multi-thresholding problem, the above equation can be written as

$$H(T) = h_0(t_1) + h_1(t_2), \ldots, h_{L-1}(t_{k-1}) \tag{2}$$

Finally, the optimal threshold ($T*$) represented in Eq. 3 can be arrived by maximizing the Shannon's entropy value.

$$T^* = \max_T\{H(T)\} \tag{3}$$

Equations (1), (2), and (3) shows image value, thresholds and cost value, respectively.

In this paper, the optimal $T*$ is discovered with the help of heuristic algorithm. For RGB image, *separate* $T*$ values are attained for R/G/B channels.

2.2 Social Group Optimization

SGO is a recently proposed heuristic approach by Satapathy and Naik to solve the constrained and unconstrained engineering optimization problem [17]. SGO consists of the following stages: (i) Educating stage to organize the position of citizens with respect to the cost function and (ii) Knowledge accomplishing phase to allow the citizens to discover optimal results.

The arithmetical form of SGO is shown in Eqs. (4)–(6)

Let

$$G_{\text{finest}} = \text{maximum}\{f(H_v)\,\text{for}\, v = 1, 2, \ldots N\} \tag{4}$$

where H_v denotes the preliminary data the people have, v =1, 2, 3,..., N specify total strength in group, and f_w is the fitness value.

The civilizing stage will alter the orientation of the agents as given below

$$H_{\text{updated}_{v,w}} = c * H_{\text{initial}_{v,w}} + R * (G_{\text{finest}_w} - X_{\text{initial}_{v,w}}) \tag{5}$$

here, H_{updated} represents renewed location, H_{initial} is the early location, G_{finest} specifies global position, R is random numeral [0, 1], and c represents the self-introspection value with a choice [0, 1]. The constant c assigned with 0.2 [18].

At the second stage, citizens are motivated to achieve the global position as follows:

$$H_{\text{updated}\,v,w} = X_{\text{initial}_{v,w}} + r1 * (H_{v,w} - H_{R,w}) + r2 * (G_{\text{finest}_w} - H_{v,w}) \tag{6}$$

where *r1* and *r2* represents arbitrary values of [0, 1] and $H_{R,w}$ is randomly assigned position a citizen. During the experimentation task, SGO finds the optimal threshold by maximizing the Shannon's function [15].

In this paper, along with the SGO, other heuristic approaches like the Particle Swarm Optimization (PSO) [5], Bacterial Foraging Optimization (BFO) [5], Firefly Algorithm (FA) [7], and Bat Algorithm (BA) [2, 3] are also considered to find the $T*$ value for the test image.

2.3 Performance Measures

In this work, a novel cost function ($J_{\max}$) presented in Eq. (7) is proposed to enhance the quality of the thresholded RGB picture.

$$J\text{max} = w_1 \times J\text{max}(t) + w_2 \times \text{PSNR} + w_3 \times \text{SSIM} \tag{7}$$

where w_1, w_2*, and* w_3 are weighting functions assigned as w_1 =1, w_2 = 0.5, and w_3 = 0.5 [7].

After the preprocessing procedure with the chosen heuristic algorithms with Shannon's entropy, the quality of output image is evaluated with a relative analysis between the original and the thresholded picture and the well known image quality measures such as MSC, PSNR, SSIM, NCC, AD, and SC are computed as discussed in [3].

Based on these values, the performance of the heuristic algorithms is evaluated [19–21].

3 Results and Discussions

The proposed image preprocessing experiment is implemented using the Matlab software. The initial algorithm parameters are assigned as follows; number of agents is chosen as 30, dimension of the search is chosen as required thresholds "*T*", total iteration value is fixed as 1500 and the stopping criterion for the heuristic search is set as $J_{\max}$. In this paper benchmark RGB images like, Barbara (720 × 576), Gold hill (720 × 576), Lena (512 × 512), Jet (512 × 612), Butterfly (481 × 321), and Starfish (481 × 321) are considered for the study. Figure 2 depicts the chosen image frames and its RGB histograms. From this, it can be noted that the RGB histogram is complex and finding an optimal thresholding is difficult with the traditional approaches.

Initially, the SGO + Shannon based thresholding is implemented on the considered RGB picture with $T = 2$, 3, 4 and 5 and the corresponding results are tabulated. The experimental work of this study confirms that computational time increases exponentially with an increase in the chosen threshold values. Table 1 presents the thresholded RGB images. Later, a relative analysis between the original and the thresholded image is implemented and the corresponding image quality measures are recorded as in Table 2. The average value of this table confirms that proposed CF offers better values of PSNR and SSIM. A similar procedure is repeated with other heuristic algorithms, like PSO, BFO, FA, and BA and the corresponding results are noted.

In order to confirm the superiority of the SGO-assisted approach, the average values of PSNR and the SSIM is then compared with the PSO, BFO, FA, and BA and the corresponding results are presented in Fig. 3a and b, respectively. From this result, it is confirmed that, SGO-based approach offers better PSNR and SSIM values compared with the other heuristic algorithms considered in this paper.

This study also confirms that the CPU run time taken by the SGO is smaller compared with the PSO and BFO approaches. The number of initial algorithm parameters to be assigned is also lesser compared with the FA and BA. Hence, in future, the SGO-assisted image processing approach can be considered to examine the grayscale and RGB images.

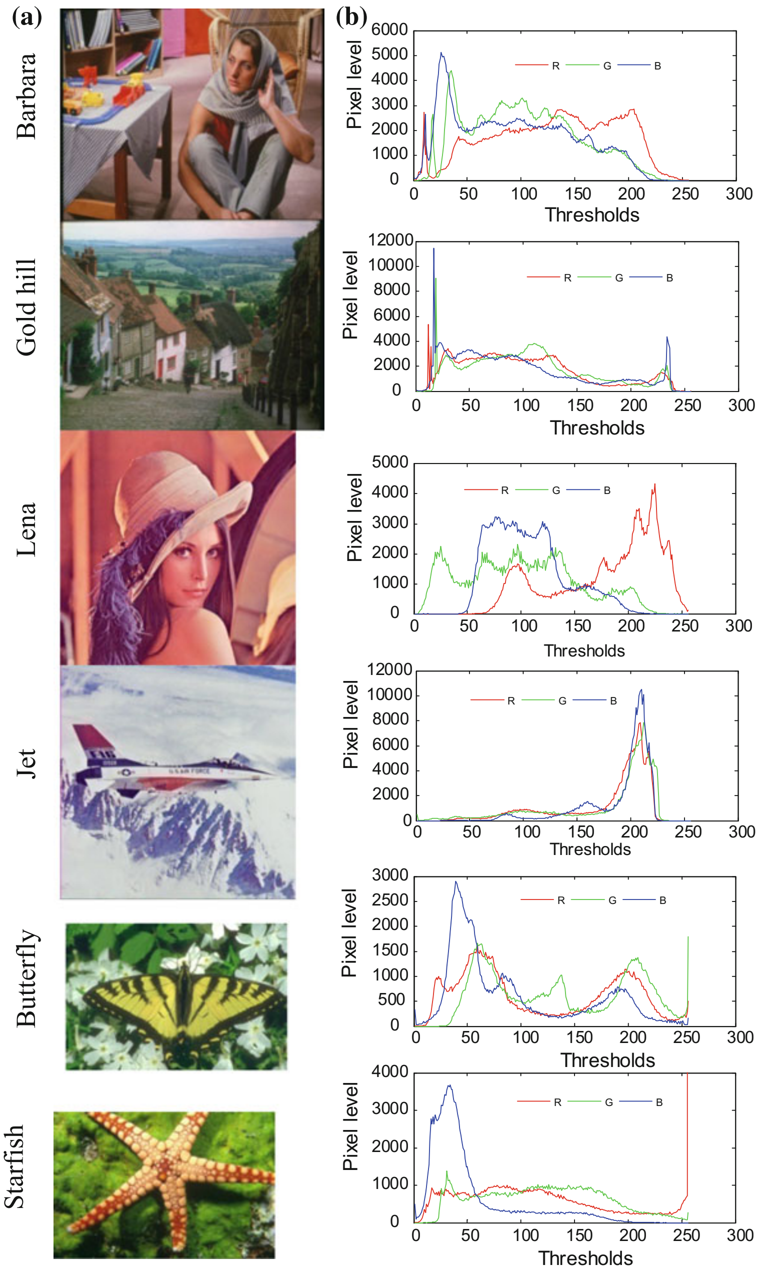
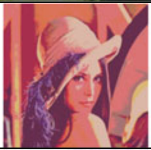
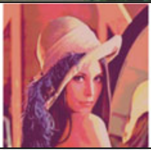
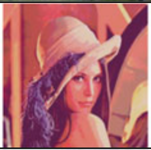
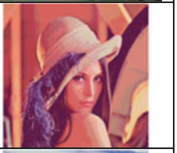
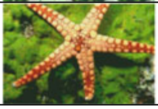
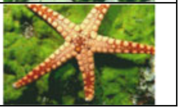
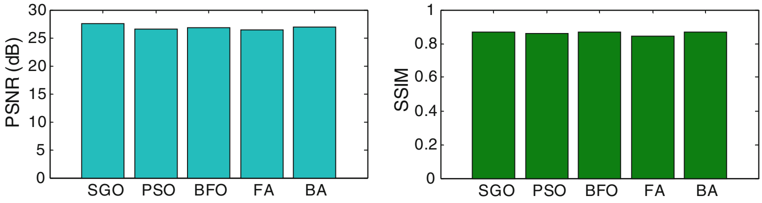


Fig. 2 RGB images considered in this paper. **a** Test image, **b** RGB histogram

Table 1 Results obtained with the SOG + Shannon

Test image	T=2	T=3	T=4	T=5
Barbara				
Gold hill				
Lena				
Jet				
Butterfly				
Starfish				



(a) Average PSNR (b) Average SSIM

Fig. 3 Validation of the considered heuristic algorithms

Table 2 Image quality measures obtained with SGO + Shannon

Image	*T*	MSE	PSNR (dB)	SSIM	NCC	AD	SC
Barbara	2	227.5507	24.5600	0.8704	0.9743	0.0364	1.0368
	3	119.0415	27.3738	0.9155	0.9831	0.1077	1.0261
	4	84.9116	28.8411	0.9352	0.9847	0.3568	1.0253
	5	59.4915	30.3863	0.9529	0.9891	0.3250	1.0179
Gold hill	2	292.5767	23.4684	0.7789	0.9621	0.3999	1.0564
	3	189.8301	25.3472	0.8114	0.9758	0.1246	1.0351
	4	95.8298	28.3158	0.8936	0.9852	0.2192	1.0227
	5	71.1092	29.6115	0.9064	0.9885	0.1496	1.0179
Lena	2	203.8198	25.0383	0.8359	0.9815	0.0303	1.0262
	3	96.2715	28.2958	0.8908	0.9894	0.0255	1.0160
	4	74.2681	29.4228	0.9072	0.9890	0.3312	1.0182
	5	43.3858	31.7573	0.9377	0.9936	0.1068	1.0104
Jet	2	195.9939	25.2084	0.8816	0.9947	−0.2583	1.0048
	3	103.1888	27.9945	0.9133	0.9978	−0.2011	1.0014
	4	75.8719	29.3300	0.9321	0.9980	−0.1168	1.0017
	5	63.2052	30.1233	0.9380	1.0006	−0.4257	0.9970
Butterfly	2	231.8921	24.4779	0.7965	0.9826	−0.6297	1.0236
	3	157.3965	26.1609	0.8324	0.9848	−0.1142	1.0229
	4	87.3504	28.7182	0.8730	0.9906	0.2114	1.0146
	5	57.4689	30.5365	0.8995	0.9932	0.1418	1.0109
Starfish	2	356.0392	22.6158	0.6815	0.9609	0.1558	1.0575
	3	158.6965	26.1251	0.7875	0.9775	0.2845	1.0357
	4	90.4870	28.5649	0.8476	0.9857	−0.0113	1.0230
	5	66.3617	29.9116	0.8716	0.9886	0.0884	1.0187
Average		133.4183	27.5911	0.8704	0.9855	0.0557	1.0217

4 Conclusion

In this work, RGB image multi-thresholding approach is implemented using the SGO and Shannon's function. A novel weighted sum of cost function is also proposed to improve the PSNR and SSIM values of the preprocessed RGB images. During this procedure, the image is preprocessed with $T = 2$, 3, 4, and 5 and the efficiency of the proposed technique is assessed by computing the well-known image quality measures. The superiority of the SGO assisted thresholding is also confirmed with a comparative study using the PSO, BFO, FA, and BA algorithms existing in the literature. From this research work, it is confirmed that, SGO + Shannon's approach offers better result compared with the other algorithms considered in this paper. In the future, this technique can be considered to preprocess the grayscale and RGB images.

References

1. Mousavirad, S.J., Ebrahimpour-Komleh, H.: Multilevel image thresholding using entropy of histogram and recently developed population-based metaheuristic algorithms. Evol. Intel. **10**(1–2), 45–75 (2017)
2. Rajinikanth, V., Aashiha, J.P., Atchaya, A.: Gray-level histogram based multilevel threshold selection with bat algorithm. Int. J. Comput. Appl. **93**(16), 1–8 (2014)
3. Satapathy, S.C., Raja, N.S.M., Rajinikanth, V., Ashour, A.S. Dey, N.: Multi-level image thresholding using Otsu and chaotic bat algorithm. Neural Comput. Appl. (2016) https://doi.org/10.1007/s00521-016-2645-5
4. Rajinikanth, V., Raja, N.S.M., Satapathy, S..C: Robust color image multi-thresholding using between-class variance and cuckoo search algorithm. Adv. Intell. Syst. Comput. **433**, 379–386 (2016)
5. Manic, K.S., Priya, R.K., Rajinikanth, V.: Image multithresholding based on kapur/tsallis entropy and firefly algorithm. Indian J. Sci. Technol. **9**(12), 89949 (2016)
6. Bhandari, A.K., Kumar, A., Singh, G.K.: Modified artificial bee colony based computationally efficient multilevel thresholding for satellite image segmentation using Kapur's, Otsu and Tsallis functions. Expert Syst. Appli. **42**, 1573–1601 (2015)
7. Rajinikanth, V., Couceiro, M.S.: Optimal multilevel image threshold selection using a novel objective function. Adv. Intell. Syst. Comput. **340**, 177–186 (2015)
8. Kumar, R., Rajan, A., Talukdar, F.A., et al.: Neural. Comput. Appl. (2016). https://doi.org/10.1007/s00521-016-2267-y
9. Li, Z., Dey, N., Ashour, A.S., Cao, L., Wang, Y., Wang, D., McCauley, P., Balas, V.E., Kai Shi, K., Shi, F.: Convolutional neural network based clustering and manifold learning method for diabetic plantar pressure imaging dataset. J. Med. Imaging and Health Informatics **7**(3), 639–652 (2017)
10. Ngan, T.T., Tuan, T.M., Minh, N.H., Dey, N.: Decision making based on fuzzy aggregation operators for medical diagnosis from dental x-ray images. J. Med. Syst. **40**(12), 280 (2016)
11. Raja, N.S.M., Rajinikanth, V., Fernandes, S.L., Satapathy, S.C.: Segmentation of breast thermal images using Kapur's entropy and hidden Markov random field. J. Med. Imaging Health Informatics **7**(8), 1825–1829 (2017)
12. Rajinikanth, V., Raja, N.S.M., Satapathy, S.C., Fernandes, S.L.: Otsu's multi-thresholding and active contour snake model to segment dermoscopy images. J. Med. Imaging Health Informatics **7**(8), 1837–1840 (2017)

13. Rajinikanth, V., Raja, N.S.M., Kamalanand, K.: Firefly algorithm assisted segmentation of tumor from brain MRI using Tsallis function and Markov random field. J. Control Eng. Appl. Inform. **19**(3), 97–106 (2017)
14. Rajinikanth, V., Couceiro, M.S.: Multilevel segmentation of color image using Lévy driven BFO algorithm. In: Proceedings of ICONIAAC14, Article No. 19, 2014. https://doi.org/10.1145/2660925
15. Kannappan, P.L.: On Shannon's entropy, directed divergence and inaccuracy. Probab. Theory Rel. Fields. **22**, 95–100 (1972). https://doi.org/10.1016/S0019-9958(73)90246-5
16. Paul, S., Bandyopadhyay, B.: A novel approach for image compression based on multi-level image thresholding using Shannon entropy and differential evolution. In: Students' Technology Symposium (TechSym), IEEE. pp. 56–61 (2014). https://doi.org/10.1109/techsym.2014.6807914
17. Satapathy, S., Naik, A.: Social group optimization (SGO): a new population evolutionary optimization technique. Complex Intell. Syst. **2**(3), 173–203 (2016)
18. Naik, A., Satapathy, S.C., Ashour, A.S., Dey, N.: Social group optimization for global optimization of multimodal functions and data clustering problems. Neural Comput. Appl. (2016). https://doi.org/10.1007/s00521-016-2686-9
19. Bhateja, V., Tripathi, A., Sharma, A., Le, B.N., Satapathy, S.C., Nguyen, G.N., Le, D.-N.: Ant colony optimization based anisotropic diffusion approach for despeckling of SAR images. Lect. Notes Comput. Sci. **9978**, 389–396 (2016). https://doi.org/10.1007/978-3-319-49046-5_33
20. Gupta, P., Srivastava, P., Bhardwaj, S., Bhateja, V.: A modified PSNR metric based on HVS for quality assessment of color images. In: International Conference on Communication and Industrial Application (ICCIA), IEEE (2011). https://doi.org/10.1109/iccinda.2011.6146669
21. Anitha, P., Bindhiya, S., Abinaya, A., Satapathy, S.C., Dey, N., Rajinikanth, V.: RGB image multi-thresholding based on Kapur's entropy—a study with heuristic algorithms. In: Second International Conference on Electrical, Computer and Communication Technologies (ICECCT), IEEE (2017). https://doi.org/10.1109/icecct.2017.8117823

Effective Neural Solution for Multi-criteria Word Segmentation

Han He, Lei Wu, Hua Yan, Zhimin Gao, Yi Feng and George Townsend

Abstract We present a novel and elegant deep learning solution to train a single joint model on multi-criteria corpora for Chinese Word Segmentation (CWS) challenge. Our innovative design requires no private layers in model architecture, instead, introduces two artificial tokens at the beginning and ending of input sentence to specify the required target criteria. The rest of the model including Long Short-Term Memory (LSTM) layer and Conditional Random Fields (CRFs) layer remains unchanged and is shared across all datasets, keeping the size of parameter collection minimal and constant. On Bakeoff 2005 and Bakeoff 2008 datasets, our innovative design has surpassed the previous multi-criteria learning results. Testing results on two out of four datasets even have surpassed the latest state-of-the-art single-criterion learning scores. To the best knowledge, our design is the first one that has achieved the latest state-of-the-art performance on such large-scale datasets. Source codes and corpora of this paper are available on GitHub (https://github.com/hankcs/multi-criteria-cws).

H. He · L. Wu (✉) · H. Yan
Computer and Software Engineering, Institutional Research,
2700 Bay Area Blvd., Houston, TX 77058, USA
e-mail: wul@uhcl.edu

H. He
e-mail: heh1996@uhcl.edu

H. Yan
e-mail: yan@uhcl.edu

Z. Gao
Computer Science Department, 3551 Cullen Blvd., Houston, TX 77204, USA
e-mail: zgao5@uh.edu

Y. Feng · G. Townsend
Department of Computer Science, Algoma University,
1520 Queen Street East, Sault Ste., Marie, ON P6A 2G4, Canada
e-mail: feng@algomau.ca

G. Townsend
e-mail: townsend@algomau.ca

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_14

1 Introduction

Unlike English language with space between every word, Chinese language has no explicit word delimiters. Therefore, Chinese Word Segmentation (CWS) is a preliminary preprocessing step for Chinese language processing tasks. Following Xue [1], most approaches consider this task as a sequence tagging task, and solve it with supervised learning models such as Maximum Entropy (ME) [2] and Conditional Random Fields (CRFs) [3, 4]. These early models require heavy handcrafted feature engineering within a fixed size window.

With the rapid development of deep learning, neural network word segmentation approach arose to reduce efforts in feature engineering [5–10]. Zheng et al. [5] replaced raw character with its embedding as input, adapted the sliding-window-based sequence labeling [6]. Pei et al. [7] extended Zheng et al. [5]'s work by exploiting tag embedding and bigram features. Chen et al. [8] employed LSTM to capture long-distance preceding context.

Novel algorithms and deep models are not omnipotent. Large-scale corpus is also important for an accurate CWS system. Although there are many segmentation corpora, these datasets are annotated in different criteria, making it hard to fully exploit these corpora, which are shown in Table 1.

Recently, Chen et al. [11] designed an adversarial multi-criteria learning framework for CWS. However, their models have several complex architectures, and are not comparable with the state-of-the-art results.

In this paper, we propose a smoothly jointed multi-criteria learning solution for CWS by adding two artificial tokens at the beginning and ending of input sentence to specify the required target criteria. We have conducted various experiments on 8 segmentation criteria corpora from SIGHAN Bakeoff 2005 and 2008. Our models improve performance by transferring learning on heterogeneous corpora. The final scores have surpassed previous multi-criteria learning, two out of four even have surpassed previous preprocessing heavy state-of-the-art single-criterion learning results.

The contributions of this paper could be summarized as:

- Proposed an effective yet elegant deep learning solution to perform multi-criteria learning on multiple heterogeneous segmentation criteria corpora;
- Two out of four datasets have surpassed the state-of-the-art scores on Bakeoff 2005;
- Extensive experiments on up to 8 datasets have shown that our novel deep learning solution has significantly improved the performance.

Table 1 Illustration of different segmentation criteria on SIGHAN bakeoff 2005

Corpora	Li	Le	reaches	Coca-Cola	Inc
pku	李	乐	到达	可口可乐	公司
msr	李乐		到达	可口可乐公司	
as	李樂		到達	可口可樂	公司
cityu	李樂		到達	可口可樂	公司

2 Related Work

In this section, we review the previous works from two directions, which are Chinese Word Segmentation and multi-task learning.

2.1 *Chinese Word Segmentation*

Chinese Word Segmentation has been a well-studied problem for decades [12]. After pioneer Xue [1] transformed CWS into a character-based tagging problem, Peng et al. [4] adopted CRF as the sequence labeling model and showed its effectiveness. Following these pioneers, later sequence labeling-based works [13–16] were proposed. Recent neural models [5, 7, 8, 11, 17] also followed this sequence labeling fashion.

2.2 *Multi-task Learning*

Compared to single-task learning, multi-task learning is relatively harder due to the divergence between tasks and heterogeneous annotation datasets. Recent works have started to explore joint learning on Chinese word segmentation or part-of-speech tagging [18, 19]. Chen et al. [11] designed a complex framework involving sharing layers with a Generative Adversarial Nets (GANs) to extract the criteria-invariant features and dataset related private layers to detect criteria-related features. This research work didn't show great advantage over previous state-of-the-art single-criterion learning scores.

Our solution is greatly motivated by Google's Multilingual Neural Machine Translation System. Johnson et al. [20] proposed an extremely simple solution without any complex architectures or private layers. They added an artificial token corresponding to parallel corpora and train them jointly, which inspired our design.

3 Neural Architectures for Chinese Word Segmentation

A prevailing approach to Chinese Word Segmentation is casting it to character-based sequence tagging problem [1, 16]. One commonly used tagging set is $\mathcal{T} = \{B, M, E, S\}$, representing the **b**egin, **m**iddle, **e**nd of a word, or single character forming a word. Given a sequence $\mathbf{X}$ with n characters as $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$, sequence tagging based CWS is to find the most possible tags $\mathbf{Y}^* = \{\mathbf{y}_1^*, \ldots, \mathbf{y}_n^*\}$:

$$\mathbf{Y}^* = \underset{\mathbf{Y} \in \mathcal{T}^n}{\arg\max}\, p(\mathbf{Y}|\mathbf{X}), \tag{1}$$

where $\mathcal{T} = \{B, M, E, S\}$. We model them jointly using a conditional random field, mostly following the architecture proposed by Lample et al. [21], via stacking Long Short-term Memory Networks (LSTMs) [22] with a CRFs layer on top of them.

We will introduce our neural framework bottom-up. The bottom layer is a character Bi-LSTM (bidirectional Long Short-Term Memory Network) [23] taking character embeddings as input, outputs each character's contextual feature representation:

$$\mathbf{h}_t = \text{BiLSTM}(\mathbf{X}, t) \tag{2}$$

After a contextual representation $\mathbf{h}_t$ is generated, it will be decoded to make a final segmentation decision. We employed a Conditional Random Fields (CRF) [3] layer as the inference layer. CRF inference layer produces global scores which is normalized to a probability in Eq. (1) via a softmax overall possible tag sequences:

$$p(\mathbf{Y}|\mathbf{X}) = \frac{e^{\text{score}(\mathbf{X},\mathbf{Y})}}{\sum_{\widetilde{\mathbf{Y}}\in\mathbf{Y}_{\mathbf{X}}} e^{\text{score}(\mathbf{X},\widetilde{\mathbf{Y}})}} \tag{3}$$

In decoding phase, first-order linear chain CRFs only model bigram interactions between output tags, so the maximum of a posteriori sequence $\mathbf{Y}^*$ in Eq. 1 can be computed using dynamic programming.

4 Elegant Solution for Multi-criteria Chinese Word Segmentation

For closely related multiple task learning like multilingual translation system, Johnson et al. [20] proposed a simple and practical solution. It only needs to add an artificial token at the beginning of the input sentence to specify the required target language, no need to design complex private encoder–decoder structures.

We follow their spirit and add two artificial tokens at the beginning and ending of input sentence to specify the required target criteria. For instance, sentences in SIGHAN Bakeoff 2005 will be designed to have the following form (Table 2):

Table 2 Illustration of adding artificial tokens into four datasets on SIGHAN Bakeoff 2005. To be fair, these <dataset> and </dataset> tokens will be removed when computing scores

Corpora	Li Le reaches Coca-Cola Inc
pku	<pku> 李 乐 到达 可口可乐 公司 </pku>
msr	<msr> 李乐 到达 可口可乐公司 </msr>
as	<as> 李樂 到達 可口可樂 公司 </as>
cityu	<cityu> 李樂 到達 可口可樂 公司 </cityu>

These artificial tokens specify which dataset the sentence comes from. They are treated as normal tokens, or more specifically, a normal character. With their help, instances from different datasets can be seamlessly put together and jointly trained, without extra efforts. These two special tokens are designed to carry criteria- related information across long dependencies, affecting the context representation of every character, and finally to produce segmentation decisions matching target criteria. At test time, those tokens are used to specify the required segmentation criteria.

5 Training

The training procedure is to maximize the log-probability of the gold tag sequence:

$$\log(p(\mathbf{Y}|\mathbf{X})) = \mathrm{score}(\mathbf{X}, \mathbf{Y}) - \operatorname*{logadd}_{\widetilde{\mathbf{Y}} \in \mathbf{Y}_{\mathbf{X}}} \mathrm{score}(\mathbf{X}, \widetilde{\mathbf{Y}}), \tag{4}$$

where $\mathbf{Y}_{\mathbf{X}}$ represents all possible tag sequences for a sentence $\mathbf{X}$.

6 Experiments

We conducted various experiments to verify the following questions:

1. Is our multi-criteria solution capable of learning heterogeneous datasets?
2. Can our solution be applied to large-scale corpus groups consisting of tiny and informal texts?
3. More data, better performance?

Our implementation is based on Dynet [24], a dynamic neural net framework for deep learning. Additionally, we implement the CRF layer in Python, and integrated the official score script to verify our scores.

6.1 Datasets

To explore the first question, we have experimented on the four prevalent CWS datasets from SIGHAN2005 [25] as these datasets are commonly used by previous state-of-the-art research works. To challenge questions 2 and 3, we applied our solution on SIGHAN2008 dataset [26], which is used to compare our approach with other state-of-the-art multi-criteria learning works under a larger scale data size. Specially, the Traditional Chinese corpora CityU, AS and CKIP are converted to Simplified Chinese using the popular Chinese NLP tool HanLP.[1]

[1] https://github.com/hankcs/HanLP.

Table 3 Comparison with previous state-of-the-art models of results on all four Bakeoff-2005 datasets. Results with ♣ used external dictionary or corpus, with ♠ are from Cai et al. [9]'s runs on their released implementations without dictionary, with ◇ expurgated long words in test set

Models	PKU	MSR	CityU	AS
Tseng et al. [13]	95.0	96.4	–	–
Zhang and Clark [27]	95.0	96.4	–	–
Zhao and Kit [28]	95.4	**97.6**	96.1	**95.7**
Sun et al. [29]	95.2	97.3	–	–
Sun et al. [16]	95.4	97.4	–	–
Zhang et al. [30]♣	96.1	97.4	–	–
Chen et al. [31]♠	94.5	95.4	–	–
Chen et al. [8]♠	94.8	95.6	–	–
Chen et al. [11]	94.3	96.0	–	94.8
Cai et al. [10]◇	95.8	97.1	95.6	95.3
Wang et al. [32]	95.7	97.3	–	–
baseline	95.2	97.3	95.1	94.9
+multi	**95.9**	97.4	**96.2**	95.4

All experiments are conducted with official Bakeoff scoring program.[2] calculating precision, recall, and F_1-score.

6.2 Results on SIGHAN Bakeoff 2005

Our baseline model is Bi-LSTM-CRFs trained on each datasets separately. Then we improved it with multiple criteria learning. The final F_1 scores are shown in Table 3.

According to this table, we find that multi-criteria learning boosts performance on every single dataset. Compared to single-criterion learning models (baseline), multi-criteria learning model (+multi) outperforms all of them by up to 1.1%. Our joint model does not rob performance from one dataset to pay another, but share knowledge across datasets and improve performance on all datasets.

6.3 Results on SIGHAN Bakeoff 2008

SIGHAN bakeoff 2008 [26] provided as many as 5 heterogeneous corpora. With another 3 non-repetitive corpora from SIGHAN bakeoff 2005, they form another standard dataset for multi-criteria CWS benchmark. We repeated our experiments on these 8 corpora and compared our results with state-of-the-art scores, as listed in Table 4.

[2] http://www.sighan.org/bakeoff2003/score This script rounds a score to one decimal place.

Table 4 Results on test sets of eight standard CWS datasets. Here, P, R, and F indicates the precision, recall, F_1 value, respectively. The maximum F_1 values are highlighted for each dataset

Models		MSR	AS	PKU	CTB	CKIP	CITYU	NCC	SXU	Avg.
Single-criterion learning										
Cheng et al. [11]	P	95.70	93.64	93.67	95.19	92.44	94.00	91.86	95.11	93.95
	R	95.99	94.77	92.93	95.42	93.69	94.15	92.47	95.23	94.33
	F	95.84	94.20	93.30	95.30	93.06	94.07	92.17	95.17	94.14
Ours	P	97.17	95.28	94.78	95.14	94.55	94.86	93.43	95.75	95.12
	R	97.40	94.53	95.66	95.28	93.76	94.16	93.74	95.80	95.04
	F	97.29	94.90	95.22	95.21	94.15	94.51	93.58	95.78	95.08
Multi-criteria learning										
Cheng et al. [11]	P	95.95	94.17	94.86	96.02	93.82	95.39	92.46	96.07	94.84
	R	96.14	95.11	93.78	96.33	94.70	95.70	93.19	96.01	95.12
	F	96.04	94.64	94.32	**96.18**	94.26	95.55	92.83	96.04	94.98
Ours	P	97.38	96.01	95.37	95.69	96.21	95.78	94.26	96.54	95.82
	R	97.32	94.94	96.19	96.00	95.27	95.43	94.42	96.44	95.64
	F	**97.35**	**95.47**	**95.78**	95.84	**95.73**	**95.60**	**94.34**	**96.49**	**95.73**

In the first block for single-criterion learning, we can see that our implementation is generally more effective than Cheng et al. [11]'s. In the second block for multi-criteria learning, this disparity becomes even significant. And we further verified that every dataset benefit from our joint-learning solution. We also find that more data, even annotated with different standards or from different domains, brings better performance. Almost every dataset benefits from the larger scale of data. In comparison with large datasets, tiny datasets gain more performance growth.

7 Conclusions and Future Works

7.1 Conclusions

In this paper, we have presented a practical way to train multiple- criteria CWS model. This effective and elegant machine learning solution only needs adding two artificial tokens at the beginning and ending of input sentence to specify the required target criteria. All the rest of model architectures, hyperparameters, parameters and feature space are shared across all datasets. Experiments showed that our multi-criteria model can transfer knowledge between differently annotated corpora. Our system is highly end-to-end, capable of learning large-scale datasets, and outperforms the latest state-of-the-art multi-criteria CWS works.

7.2 Future Works

Our highly effective and elegant multi-criteria learning technique can be applied to any sequence labeling task such as POS tagging and NER. We plan to conduct more experiments of using our novel effective machine learning technique in various application domains.

References

1. Xue, N.: Chinese word segmentation as character tagging. IJCLCLP (2003)
2. Jin, K.L., Ng, H.T., Guo, W.: A maximum entropy approach to chinese word segmentation. In: Proceedings of the Fourth Sighan Workshop on Chinese Language Processing (2005)
3. Lafferty, J.D., Mccallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Eighteenth International Conference on Machine Learning, pp. 282–289 (2001)
4. Peng, F., Feng, F., Mccallum, A.: Chinese segmentation and new word detection using conditional random fields, pp. 562–568 (2004)
5. Zheng, X., Chen, H., Xu, T.: Deep learning for Chinese word segmentation and POS tagging. EMNLP (2013)

6. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.P.: Natural language processing (Almost) from scratch. J. Machine Learning Res. (2011)
7. Pei, W., Ge, T., Chang, B.: Max-margin tensor neural network for Chinese word segmentation. ACL (2014)
8. Chen, X., Qiu, X., Zhu, C., Liu, P., Huang, X.: Long short-term memory neural networks for Chinese word segmentation. EMNLP (2015)
9. Cai, D., Zhao, H.: Neural word segmentation learning for Chinese. ACL (2016)
10. Cai, D., Zhao, H., Zhang, Z., Xin, Y., Wu, Y., Huang, F.: Fast and Accurate Neural Word Segmentation for Chinese (April 2017). arXiv:1704.07047
11. Chen, X., Shi, Z., Qiu, X., Huang, X.: Adversarial multi-criteria learning for Chinese word segmentation. **1704** (2017). arXiv:1704.07556
12. Huang, C., Zhao, H.: Chinese word segmentation: a decade review. J. Chin. Inf. Process. **21**(3), 8–19 (2007)
13. Tseng, H., Chang, P., Andrew, G., Jurafsky, D., Manning, C.: A conditional random field word segmenter for sighan bakeoff **2005**, 168–171 (2005)
14. Zhao, H., Huang, C., Li, M., Lu, B.L.: Effective tag set selection in Chinese word segmentation via conditional random field modeling. PACLIC (2006)
15. Zhao, H., Huang, C.N., Li, M., Lu, B.L.: A unified character-based tagging framework for chinese word segmentation. Acm Trans. Asian Language Inf. Process. **9**(2), 1–32 (2010)
16. Sun, X., Wang, H., Li, W.: Fast online training with frequency-adaptive learning rates for chinese word segmentation and new word detection, pp. 253–262 (2012)
17. Dong, C., Zhang, J., Zong, C., Hattori, M., Di, H.: Character-based LSTM-CRF with radical-level features for Chinese named entity recognition. NLPCC/ICCPOL (2016)
18. Li, Z., Chao, J., Zhang, M., Chen, W.: Coupled Sequence Labeling on Heterogeneous Annotations: POS Tagging as a Case Study. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Stroudsburg, PA, USA, Association for Computational Linguistics, pp. 1783–1792 (2015)
19. Chao, J., Li, Z., Chen, W., Zhang, M.: Exploiting heterogeneous annotations for Weibo word segmentation and POS tagging. NLPCC (2015)
20. Johnson, M., Schuster, M., Le, Q.V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F.B., Wattenberg, M., Corrado, G., Hughes, M., Dean, J.: Google's Multilingual Neural Machine Translation System - Enabling Zero-Shot Translation. **cs.CL** (2016)
21. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. CoRR (2016)
22. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
23. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks **18**(5–6), 602–610 (2005)
24. Neubig, G., Dyer, C., Goldberg, Y., Matthews, A., Ammar, W., Anastasopoulos, A., Ballesteros, M., Chiang, D., Clothiaux, D., Cohn, T., et al.: Dynet: the dynamic neural network toolkit. (2017) arXiv preprint arXiv:1701.03980
25. Emerson, T.: The second international chinese word segmentation bakeoff. In: Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, Jeju Island, Korea, pp. 123–133 (2005)
26. MOE, P.: The fourth international chinese language processing bakeoff: Chinese word segmentation, named entity recognition and chinese pos tagging. In: Proceedings of the sixth SIGHAN workshop on Chinese language processing (2008)
27. Zhang, Y., Clark, S.: Chinese segmentation with a word-based perceptron algorithm. In: Czech Republic, Association for Computational Linguistics, pp. 840–847. Prague (2007)
28. Zhao, H., Kit, C.: Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition. In: The Sixth SIGHAN Workshop on Chinese Language Processing, pp. 106–111 (2008)

29. Sun, X., Zhang, Y., Matsuzaki, T., Tsuruoka, Y., Tsujii, J.: A discriminative latent variable chinese segmenter with hybrid word/character information, pp. 56–64 (2009)
30. Zhang, L., Wang, H., Sun, X., Mansur, M.: Exploring representations from unlabeled data with co-training for chinese word segmentation. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Washington, USA, Association for Computational Linguistics, pp. 311–321 (October 2013)
31. Chen, X., Qiu, X., Zhu, C., Huang, X.: Gated recursive neural network for Chinese word segmentation. ACL (2015)
32. Wang, C., Xu, B.: Convolutional neural network with word embeddings for chinese word segmentation. (2017). arXiv preprint arXiv:1711.04411

Detection and Classification of Trendy Topics for Recommendation Based on Twitter Data on Different Genre

D. N. V. S. L. S. Indira, R. Kiran Kumar, G. V. S. N. R. V. Prasad and R. Usha Rani

Abstract Microblogging encourages you to impart or depict your present status in the form of short posts by making use of instant messaging, emails and social media. At the same time, this is continuously generating huge amounts of raw un-structured data which has proved to be a complex task for analysis. In this paper, a solution based on classification of selected features is proposed to address the problem. In the proposed work, we considered limited number of pre-defined generic classes such as sports, politics, entertainment, electronics, and others. To achieve this, a novel algorithm is proposed, which performs re-iterative clustering on the trained data including the new subset of tweets added. Once the tweets are classified, we then proceed to find the genre to which the trending topic belongs. Combining this classification with the polarity retrieval, algorithms will help for recommending a topic of a given genre on a given days' tweets.

1 Introduction

In the recent times, as the growth of internet and communication over mobile phones has increased, short text has become a new appearance of text. Short texts such as the snippets, Question and Answer passages, image captions, and product descriptions

D. N. V. S. L. S. Indira (✉) · G. V. S. N. R. V. Prasad
Department of Computer Science and Engineering, Gudlavalleru Engineering College (Affiliated to Jawaharlal Nehru Technological University Kakinada), Gudlavalleru, Krishna District, AP, India
e-mail: indiragamini@gmail.com

G. V. S. N. R. V. Prasad
e-mail: gutta.prasad1@gmail.com

R. Kiran Kumar
Department of Computer Science, Krishna University, Machilipatnam, Krishna District, AP, India
e-mail: kirankreddi@gmail.com

R. Usha Rani
Department of Computer Science and Engineering, CVR College, Hyderabad, India
e-mail: usha.shreni@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_15

etc., have played important roles in current Web and IR applications [1]. Short text has extensively been used as a mode of text communication over phones, Microblogs etc., features of short text makes it a challenge to classify the kind of data it is. Twitter, a well-known microblogging tool has seen a great deal of development since it propelled in October 2006 [2]. Unlike normal documents short texts are noisier with length of each text message being less than 280 characters. It has only few sentences and few words which have very little effective information. It is difficult to extract accurate sample features as the dimension of feature set is very high. Next major challenge in capturing the real-time data without any data loss, most of the messages created is real-time information which constantly update in seconds and difficult to collect. These features of the short text make it difficult to obtain the required information rapidly and accurately.

These short texts help in knowing what kind of topic is trending. For example, when the most trending topic for the day is "arnab goswami" and users who are unaware of this name would not be able to make out that the name belongs to a top most journalist of India and which issue is trending highest on a given day. Hence, to make it easy for users, we define an algorithm to classify the tweets to a specific class depending on nature of the tweet and thereby recommend about the positives or negatives of the trending topics by generating polarity. In this algorithm, we addressed this problem by defining five different classes and categorized the tweets into respective class, and when a given tweet does not fall into any of the categories it goes to "others" category.

As of now, there are numerous customary text classification methods like Naive Bayes (NB), Support Vector Machines (SVM), Neural Network (NN), Decision Tree (DT), and k-Nearest Neighbor (KNN) [3], yet those techniques put long text as an examination object. To short text, the abovesaid methods have a poor performance in terms of execution because of the sparsity of features, irregularity and big data.

2 Existing Work

There have been a lot of studies that have been proposed on text classification, however, it is a very broad term to define the complete study on text classification. Existing pre-defined set of algorithms like SVM can solve the problem of data samples with high dimensions and nonlinearity. This has been applied on a lot of applications over a period on face recognition, fingerprint recognition, etc.

Similarly, there are algorithms like Naïve Bayes (NB) which is based on Bayes theory and probability research on the field of Machine Learning which can be applied on big database simply and efficiently. Apart from SVM and NB [3], there are many more algorithms like K-Nearest Neighbor, fuzzy logics, etc. These algorithms have an excellent presentation on long text classification but a poor concert on non-linear short text.

Those conventional text classification methods cannot be basically used to unravel the short text meaning and it has low ability when the span of information is substantial. So, a new method should be deliberated to solve the difficulty of finding the correct category of the tweet. This research uses a dynamic approach of finding out the features from the tweets and thereby gets clustered into appropriate category.

3 Proposed Work

A novel algorithm has been considered to evaluate and classify the tweets and thereby recommend the user about the goodness of a particular tweet or hashtag or a classified genre.

3.1 Corpus cleansing.
3.2 Tweet classification.
3.3 Polarity classification.
3.4 Recommender.

3.1 Corpus Cleansing

There are a lot of unnecessary things in the tweet which would decrease the result of classifier. Hence cleansing of the corpus is necessary before we apply our algorithm. Removal of stop words is one major aspect that helps to build a better classifier. Stop words are meaningless and have low discriminative power. We have made use of Zipf's law algorithm to avoid stop words when providing input to classifier.

George K. Zipf's law [4] says that the occurrence of a word rate in a body of text or documents is nearly opposite to the rank of that word by the number of occurrences of the most frequently occurring word

$$P_n \sim \frac{P_1}{n^{\alpha}} \tag{1}$$

—Courtesy from [4]

where P_n is the frequency of the nth ranked word, and α is close to 1.

In this study, we followed three step Zipf's law to remove stop words:

- Removing most frequent words (TF-High).
- Removing words that have occurred only once (TF1).
- Removing words with low inverse document frequency.

To implement Zipf's algorithm, we first apply TF-IDF algorithm on the complete corpus to retrieve term and document frequencies. Once the stopword list is identified using Zipf's law, it is applied on the corpus by steps given below:

A Spark Job is Implemented to Make the Below Algorithm Work as Expected

Step 1: The target document text is tokenized and individual words are stored in array.
Step 2: The stopword file is stored in cache as a Spark RDD. Every single stopword is read from the distributed cache
Step 3: The stopword is judge against to target text in form of array using chronological search procedure.
Step 4: If it matches, the word in array is detached, and the comparison is continued till length of array.
Step 5: After elimination of stopword totally, another stopword is read from stopword list and yet again algorithm follows Step 2. The algorithm runs continually until all the stopwords are compared.
Step 6: Resultant text keep away from of stopwords is displayed and collected in a file.

3.2 Tweet Classification

Classification of text data to categorize articles, books, and several page long documents are a widely used application. For a classifier, to classify the data it needs righteous information. Hence, for this purpose, the input data is split into testing and training data. However, for classification of tweet, we have very limited set of data as tweet can only be of 280 characters.

As part of tweet classification, we have considered four different genres of classifying [2, 5–9] tweets. We made use of hashtags and features extracted from trained data to cluster each genre. Genres we considered in our research are politics, electronics, animals, sports, and others.

We made use of Stanford NLP techniques in order to train the dataset, with tenfold data training algorithm to create a test data set. To achieve this, we made a feature subset for every genre with the relevant feature terms. Using this relevant feature set, data training is done to fetch related data for the feature set.

3.2.1 Data Training Algorithm

Step 1: Breakup Training data into 10 partitions with and label tweets into politics, electronics, animals, and others categories
Step 2: If the tweet contains politics tag; make a bag of words for politics with all the hashtags associated with that tweet
Step 3: For each Iteration i,

- choose one partition as test set

- Train on other nine partitions using MaxEnt classifier, compute performance on Test set
- Hence, classifier ci is generated.

Step 4: Iterate Step 2 on all the partitions

Example Let us consider a tweet and see how bag of words get created. Below is the tweet that would be an output from Sect. 3.2

Tweet1:
("M.S. Dhoni is all time best captain in the sport of cricket", ☺☺, #captain#MSD#india#cricket#champion#leader)
Now that this tweet comes into a sports genre, all the nouns (POS tagged) and the hashtags would be a bag of words for the genre SPORTS
Tweet2:
("modi has took a right decision", #namo,#primeminister#leader)
Similarly, all the above hashtags will go into genre called POLITICS

The challenging thing here is #leader is present in both the tweets and therefore would be in Sports and Politics [5] bag of words. Hence, finding the righteous genre for a given word is difficult in such scenarios, thereby once the bag of words [10] is generated using clustering techniques. In our observation, we found that 90° of the hashtags present in each tweet is used to extend or intensify or brief the tweet.

3.2.2 Algorithm: Tweet Classifier

L: Length of hashtags in H_i is L
W_i: Weight of Hashtag x_i
H_i: Bag of Hashtags

Step 1: Considering Bag of Hashtags for a given tweet T_i and H_i
Step 2: Split hashtags in H_i into comma separated values $(x_{0,}\ x_{1,}\ x_{2\,\ldots})$ based on alphabetical order.
Step 3: Assign W_i of each unique x_i to 1
Step 4: From $\mathrm{H_i}$ multiple, sub-bags are created
For $Z = 0$ to L,

$$\mathrm{Bag}_z = \sum_{i=1}^{i=L} (xi);\ W_z = Z$$

$$\mathrm{Bag}_{z+1} = \sum_{i=1}^{i=L} (xi, xi + 1);\ W_{z+1} = Z + 1$$

$$\mathrm{Bag}_{z+2} = \sum_{i=1}^{i=L} (xi, xi + 1, xI + 2);\ W_{z+2} = Z + 2$$

.

.

$$\text{Bag}_L = \sum_{i=1}^{i=L}(xi, xi + 1, xI + 2, \ldots, xL); W_i = L$$

Step 5: Cumulative weight of T_i is sum of all W_{z+1} where $\underline{i}$ =0 to L

To classify the genre of existing tweet we made use of clustering technique [10], we have considered five different genres to be classified into, and hence five different clusters are formed.

Each cluster has the below features:

- Name of the cluster.
- List of all the sub-bags of hashtags.
- Cumulative weight of cluster, which is sum of all cumulative weight of all tweets.

For every new tweet that needs to be classified, sub-bags of hashtags are generated and once all the four steps are completed for all the 100-trained dataset (tweets). Every label has a bag of words.

3.2.3 Algorithm: Genre Classifier

Step 1: For a new tweet T_i, sub-bags are generated

$$\text{For } Z = 0 \text{ to } L$$

$$\text{Bag}_z = \sum_{i=1}^{i=L}(xi); W_z = Z$$

$$\text{Bag}_{z+1} = \sum_{i=1}^{i=L}(xi, xi + 1); W_{z+1} = Z + 1$$

$$\text{Bag}_{z+2} = \sum_{i=1}^{i=L}(xi, xi + 1, xI + 2); W_{z+2} = Z + 2$$

.

$$\text{Bag}_L = \sum_{i=1}^{i=L}(xi, xi + 1, xI + 2, \ldots, xL); W_i = L$$

Step 2: Each sub-bag is compared with the bags of the clusters.

Step 3: For a multiple bag hit, heavy weighted bag is chosen as priority in classifying into a genre

3.3 Polarity Classification

To generate polarity of the short texts classified in the above section, we made use of attribute tagger algorithm [11]

3.3.1 Algorithm: Tweet Polarizer

T: Cleansed tweet
W: Tokenized words
F: Feature list
JJ: Adjective, JJR: Adjective Comparative, JJS: Adjective superlative
VNEG: Very Negative, VPOS: Very Positive
NEG: Negative, POS: Positive N: Neutral
Step 1: Tokenise T using Stanford NLP tokeniser which converts T into W
Step 2: From W features of the sentence is extracted and sent into classifier
*Feature Extraction: we use TF*IDF algorithm to calculate the weights*

$$w_{ij} = \log\left(tf_{ij}\right) * \log([m - m_i]/m_i)$$

where w_{ij} is the weight for a term i and document j, tf_{ij} is the number of times the term i occurred in document j, m is the total number of documents, and m_i is the number of documents in which term i appeared [12].

Highest weighted terms are considered as features
Step 3: Iterate F on the MaxEnt classifiers generated from Data training algorithm to find polarity based on Trained data
Step 4: Apply POS tagging on T, to retrieve JJ, JJR, JJS
if Step 3 results NEG and if the sentence contains JJ, JJR, JJS mark it as VNEG
if Step 3 results POS and if the sentence contains JJ, JJR, JJS mark it as VPOS
Step 5: convert Ratings to numeric's
VPOS to 2, POS to 1, N to 0, NEG to -1, VNEG to -2

This data in the form of CSV file
(#tag, emoticons, genre, geo points, tweet)

This data is fed into elastic search and visualized using Kibana for the end user for the data summarization.

3.4 Recommender

Once the tweets are classified and polarity is assigned to each tweet and thereby to each genre, we define a recommendation to be provided on any searched keyword based on the polarity of the subject across the users, the cumulative weight of the genre on a given day (24 h).

Table 1 Number of distinct hashtags is clustered into different genres when a new tweet comes live streamed

Feature	Value
Tweets per hour	290,080
Cleansed tweets per hour	185,000
Classified tweets into genre (not others)	18,000
Number of distinct hashtags	57,000

Table 2 Determining the time it took when the clustering had started

Nth hour	Iterations of hashtags into clusters	Time in milliseconds
1	27,000	0.2
6	79,589	7
20	189,000	20
24	215,903	23

This generates a recommender score for every given subject that is been searched for, this score is subject to change for every regular interval as the tweets gets streamed real-time into the system

$$\mathrm{RS} = \left(\sum_{0}^{n} P(i)/n(p)\right) + \left(\sum_{0}^{n} Z(i)/n(Z)\right) \tag{2}$$

RS	Recommender Score
P(i)	Polarity of user I on a given subject
Z(i)	Polarity of given genre
n(p)	Number of polarities from different users
n(Z)	Number of genres.

4 Experimental Results

To evaluate the performance of our recommendation algorithm, we have virtualized a cluster with 10 machines interconnected, with YARN resource manager and Apache Spark is installed in it. Elastic Search along with Kibana is installed to store the data generated from the spark process and visualize it real-time using Kibana (Tables 1 and 2; Figs. 1, 2, 3, and 4).

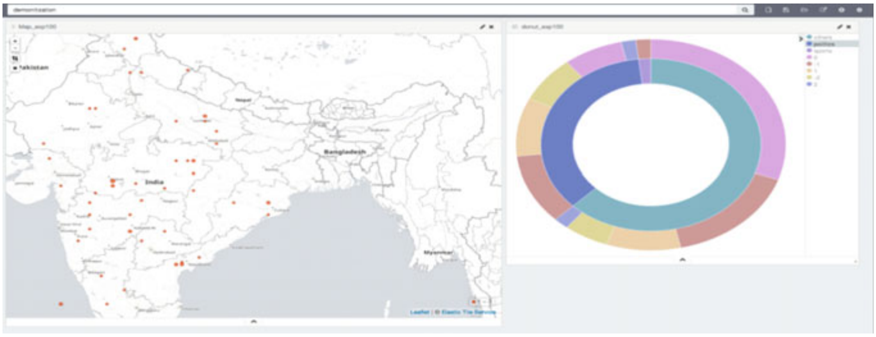
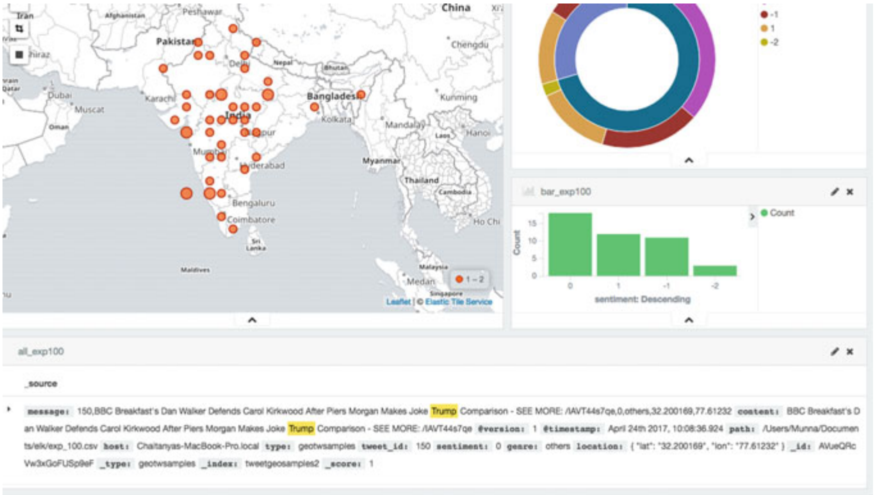


Fig. 1 In the image below, we have made a brief polarity classification on the term demonetization considering 10,000 tweets from India, which comes into genre politics as highlighted in the image below



_source

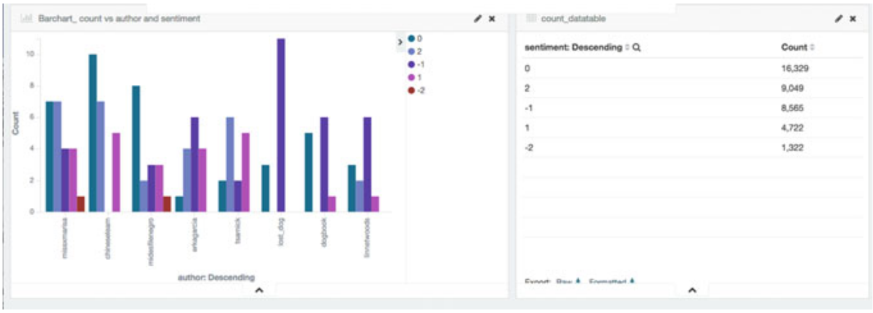
message: 150,BBC Breakfast's Dan Walker Defends Carol Kirkwood After Piers Morgan Makes Joke Trump Comparison - SEE MORE: /lAVT44s7qe,0,others,32.200169,77.61232 **content:** BBC Breakfast's Dan Walker Defends Carol Kirkwood After Piers Morgan Makes Joke Trump Comparison - SEE MORE: /lAVT44s7qe **@version:** 1 **@timestamp:** April 24th 2017, 10:08:36.924 **path:** /Users/Munna/Documents/elk/exp_100.csv **host:** Chaitanyas-MacBook-Pro.local **type:** geotwsamples **tweet_id:** 150 **sentiment:** 0 **genre:** others **location:** { "lat": "32.200169", "lon": "77.61232" } **_id:** AVueQRcVw3xGoFUSp9eF **_type:** geotwsamples **_index:** tweetgeosamples2 **_score:** 1

Fig. 2 Search analytics on term Trump along with the polarity of the subject "Trump"

5 Conclusion

The experimental results of using the proposed algorithm for classifying the raw un-structured data had yielded better results than existing algorithms in terms of recommendations made to the end user of the system. In future, this work can be extended by considering multiple corpuses for different kinds and formats of data to make every analyst flexible to bring the information out. Many researchers have done lot of analysis on text data, however, there is still scope to evolve and equip the usage with latest technologies that would give accuracy and good performance.

It is unfortunate that location-enabled tweets are currently sparse. Therefore, gathering high volume location-enabled data is a big challenge. Twitter users generally



sentiment: Descending	Count
0	16,329
2	9,049
-1	8,565
1	4,722
-2	1,322

Fig. 3 Represents the total number of tweets in *x*-axis versus sentiment, Bar chart: A stratagem of total tweet count in thousand in *y*-axis versus top eight tweeted users on *x*-axis, with a split of different attitudes

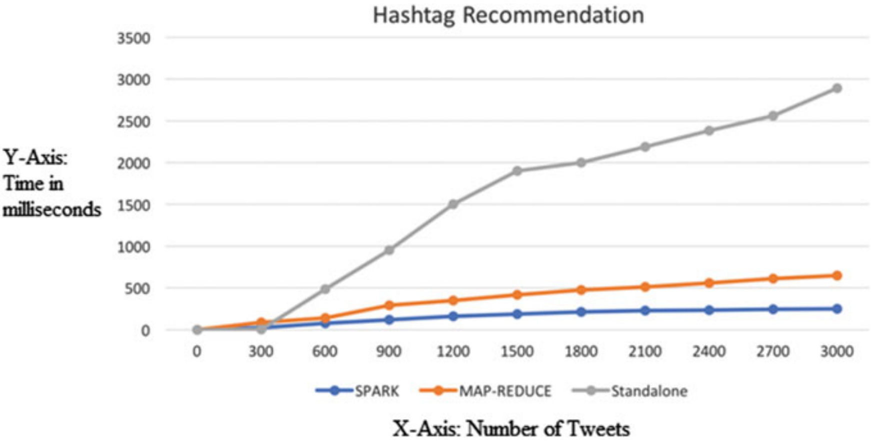


Fig. 4 Line chart describes the performance of hashtag recommender in different architectures named as standalone versus MapReduce versus spark

do not share their geo location. Twitter API also does not allow the access of location-related information such as server origin for user privacy. To overcome this problem, future work aims to develop and integrate an algorithm content-based approach to determine user's location.

References

1. Hu, X., Sun, N., Zhang, C., Chua, T.-S.: Exploiting internal and external semantics for the clustering of short texts using world knowledge. In: Proceedings CIKM, pp. 919–928. Hong Kong, China (Nov. 2009)
2. Sriram, B.: Short text classification in twitter to improve information filtering. 2010. In: Graduate Program in Computer Science and Engineering. The Ohio State University (2010)

3. Phan, X.-H., Nguyen, L.-M., Horiguchi, S: Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: Proceeding WWW, pp. 91–100 Beijing, China (Apr. 2008)
4. Chen, Y.-S.: Zipf's law in natural languages, programming languages, and command languages. In the Simon-Yule approach. Int. J. Syst. Sci. **22**(11), 2299–2312 (1991)
5. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting elections with twitter: what 140 characters reveal about political sentiment (2010)
6. Banerjee, S., Ramanathan, K., Gupta, A.: Clustering short texts using Wikipedia. In: Proceedings of 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 787–788 (2007)
7. Xiaojun, L., Meng, Z., Xiao. B.: Short text classification method based on concept network. Comput. Eng. **36**(21), 4–6 (2010)
8. Armentano, M.G., Godoy, D.L., Amandi, A.A.: Recommending information sources to information seekers in Twitter. In: International Workshop on Social Web Mining
9. Huang, J., Thornton, K.M., Efthimiadis, E.N.: Conversational tagging in Twitter. In: The 21st ACM Conference on Hypertext and Hypermedia, pp. 173–178 (2010)
10. Yu, Z., Wang, H., Lin, X., Member, S.: Understanding short texts through semantic enrichment and hashing. In: IEEE, and Min Wang **28**(2) (2016)
11. Indira, D., Kumar, R.K.: Polarity classification and intensification based on emoticons and POS tagging on Twitter Data. In: 2017 International Conference on Innovations in information Embedded and Communication Systems (ICIIECS)
12. Metzler, D., Dumais, S., Meek, C.: Similarity measures for short segments of text. Lect. Notes Comput. Sci. **4425**, 16 (2007)

An Enhanced RGB Scan Image Encryption Using Key-Based Partitioning Technique in Cloud

Karuna Arava and L. Sumalatha

Abstract Cloud is a group of remote systems available on the web to manage and store information. As a "utility" cloud provides better use of resources and minimizes initial cost. Nowaday's, importance for secured storage of Health records and Medical images has drastically improved in cloud environment. Public clouds are ubiquitous by nature, it is the user self-responsibility to achieve privacy for their own data. Existing techniques achieve confidentiality using standard encryption algorithms. This paper proposes a hybrid scheme for encryption of plain/medical images using key dynamics before publishing them. Other users can decrypt the cipher image only after they are authenticated by the owner with restricted access. Based on experimental results, a hybrid encryption scheme demonstrates excellent level of confidentiality.

1 Introduction

Many cryptographic algorithms like RSA, DES, Blowfish techniques, Hash, Chaos, Scrambling Methods, and various asymmetric and symmetric approaches exist for encryption. In 2006, Pareek, Sud, and Patidar proposed an encryption scheme based on chaotic graphs and a key with 80-bit are selected [1]. Zhou demonstrated an algorithm using binary bit plane which is generated using decomposition method [2]. Image encryption is done using symmetric chaos theory cryptosystem by applying android cat map and logistic map for bit-level transformation and scattering, respectively [3]. The difference between position-based scrambling and pixel-based scrambling are addressed based on LSB steganography [4]. Yuan and Jiang proposed image scrambling method, which reads the pixel values present in the even-numbered lines and odd-numbered lines and does scrambling to the pixel values [5]. Apart from

K. Arava (✉) · L. Sumalatha
Department of CSE, UCEK (A), JNTUK, Kakinada, India
e-mail: karunagouthana@gmail.com

L. Sumalatha
e-mail: Sumapriyatham@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_16

standard algorithms for image encryption there exist other methods like Scan pattern, where the image blocks are rearranged along with pixel values [7]. Authors proposed Cryptanalysis based on Cellular automata [8]. Panduranga proposed encryption technique using scan designs and carrier, in which image is organized with the help of alphanumeric password [9]. An image spatial acquiring technique build on formal language and scan language presents a method to parallel implementation [10]. Yue Wu and Xohu proposed ciphering using Latin squares and irregular noise is embedded into the short end significant bit plane, permutations, and combinations are applied to get secure cipher image [11, 12]. Guosheng Gu and Jie Ling proposed that the locations of the pixels are permuted and the values of the pixels are substituted at each walk of chaotic map to encrypt [13]. Studying paths for an image of pxp size established on a 2D spatial accessing technique [2].

The proposed scheme inherits the properties of two strategies (i) scan methodology and (ii) key-based partitioning technique together called as hybrid encryption scheme.

2 Preliminaries

SCAN Methodology

SCAN technique refers to various ways of scanning a two-dimensional Image. The SCAN method is an excellent way for scanning, which decreases the bit count used for ciphering and also the bit sequence. The SCAN technique is appropriate for information masking, encryption, and compression, while doing operations based on multimedia.

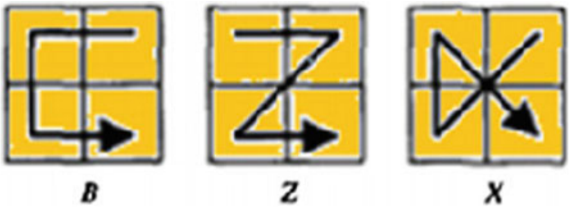
SCAN Partition Patterns

The basic patterns of partition types are represented in Fig. 1

Type 1: B-shape partition patterns.
Type 2: Z-shape partition patterns.
Type 3: X-shape partition patterns.

There is a chance to transform each type 1, 2, and 3 of basic pattern into eight different transfigurations, depending on the first and last point of the pattern. Type 1, 2, 3 patterns with its eight transfigurations can be represented from 0 to 7 in (Figs. 2, 3 and 4)

Fig. 1 Fundamental partition patterns: type 1, type 2, and type 3



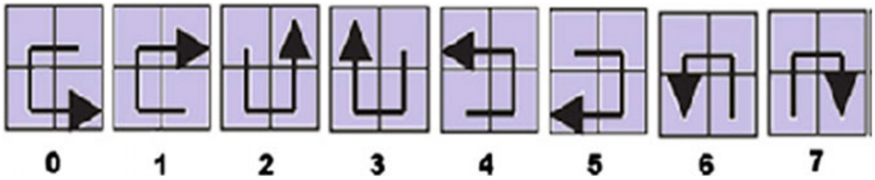


Fig. 2 Transfiguration of type 1 partition pattern from 0 to 7

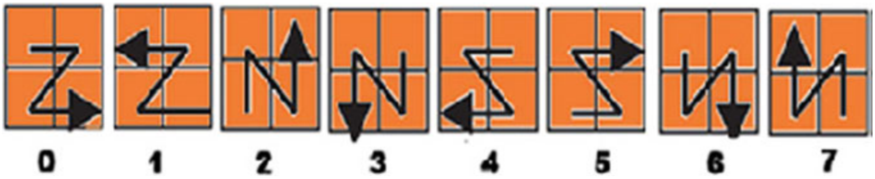


Fig. 3 Transfiguration of type 2 partition pattern from 0 to 7

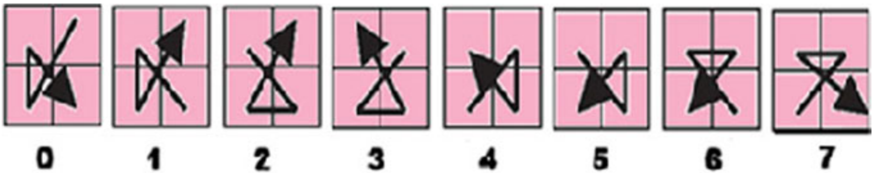


Fig. 4 Transfiguration of type 3 partition pattern from 0 to 7

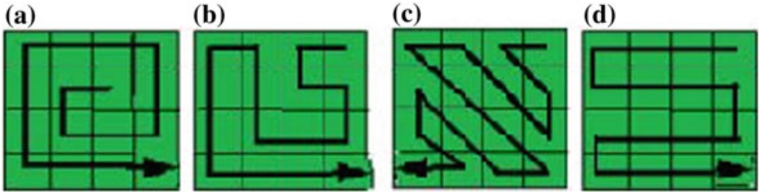


Fig. 5 Four standard scanning patterns

Standard Scanning Patterns

There are four scanning patterns described in Fig. 5 [6]

1. Coiled or Spiral Pattern (S).
2. Quadratic or Orthogonal Pattern (O).
3. Crosswise or Diagonal Pattern (D).
4. Raster Pattern (C).

The patterns of S, O, D, and C in Fig. 5a, b, c, d. Every pattern in this basic scanning pattern can be orbited through 0°, 90°, 180°, and 270° [14]. For example, the coiled patterns S0, S2, S4, and S6 with respective angles can be represented in

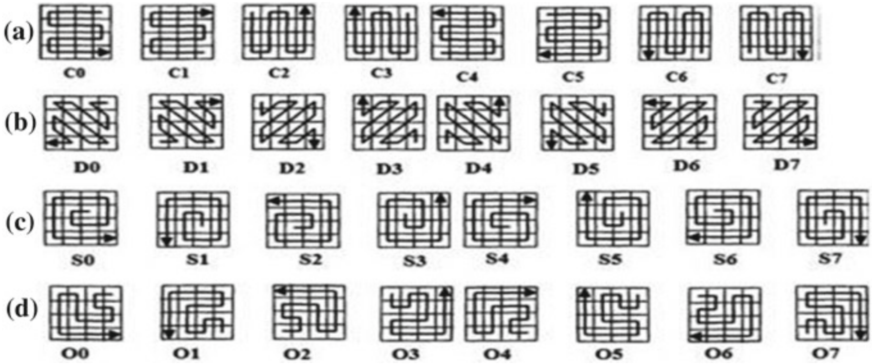


Fig. 6 The eight transfigurations of (C, D, S and O) SCAN patterns

Fig. 6c, the reverse patterns for the above patterns corresponded to the S1, S3, S5, and S7 in Fig. 6c. Similarly, the other patterns are shown in Fig. 6a, b, d.

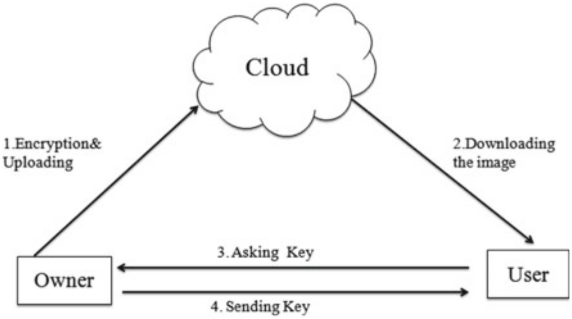
Key-Based Partitioning Technique In the key-based partitioning technique [15], the image is split into "*N*" blocks of equal size. Bit operations like Circular Right Shift Operation (CRS), Circular Left Shift Operation (CLS), NOT and Circular Right Shift Operation (NCRS). If *n* operations are chosen, then we have *n*! possibilities (say *n*=3; 1-2-3, 2-3-1, 3-2-1, 1-3-2, 2-1-3, and 3-1-2 possibilities). These patterns are applied on the blocks depending on the shift key, which is shared over a secret medium in which source and destination apply on the key image. For each block the pattern selection can be done with the help of following steps:

Step 1: Input an image I
Step 2: Split the image I into nine segments of equal size
Step 3: Number those segments as "I","II","III","IV","V","VI","VII","VIII","IX"
Step 4: Consider a Shift Key "*K*" (Say, Ex: *K* = JNTUK)
Step 5: Find *Sum* = (ASCII (J) + ASCII (N) + ASCII (T) + ASCII (U) + ASCII(K));
Step 6: Compute *value* = *Sum* % (length (k))
Step 7: *value* will be the order position in *n*! possibilities

Example: *Sum* = 74 + 78 + 84 + 85 + 75
Sum = 396
value = 396 % 5
value = 1(if 0: 1-2-3, 2-3-1, 3-2-1; if 1: 2-1-3, 3-2-1, 1-2-3 and so on)

S. No	Blocks	Pattern
1	I, II, III	2 3 1
2	IV, V, VI	3 2 1
3	VII, VIII, IX	1 3 2

Fig. 7 Architecture



3 System Architecture

In Fig. (7) in step 1, Encryption is done at owner site and Uploads encrypted image into Cloud. User can download the image from cloud and requests the key for decryption. Owner will validate user request and share the key for decryption to user for original image.

4 Proposed Algorithm

The proposed algorithm for color image encryption is based on key-based partitioning technique is explained below:

Step 1: Input an image*(X)* and Key image *(K)*

Step 2: Divide image*(X)* into *3(R, G, B)* components

$\mathrm{R} \rightarrow X\ (: ; : ; 1)$
$\mathrm{G} \rightarrow X\ (: ; : ; 2)$
$\mathrm{B} \rightarrow X\ (: ; : ; 3)$

Step 3: Apply pattern(*S*) on the components

$\mathrm{R} + S \rightarrow I1$
$\mathrm{G} + S \rightarrow I2$
$\mathrm{B} + S \rightarrow I3$

Step 4: Divide Key image *(K)* into "*N*" parts

$$P = [1, 2, 3, \ldots N]$$

Step 5: Enter key '*k*' to encrypt Key image *(K)*

Step 6: Perform *modulus (%)* operation on sum of all ASCII values of '*k*' and length of key ("l")

Step 7: Apply CRS, CLS, NCRS on "*N*" parts based on mod value

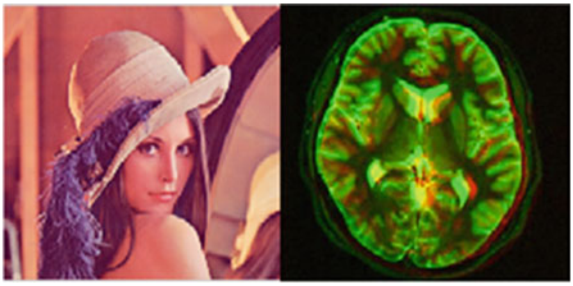


Fig. 8 Plain images: Lena and medical

Step 8: Concatenate all "*N*" parts, get image "*E*"

Step 9: Divide the image "*E*" into *3(R, G, B)* components

$E1 \rightarrow X\ (: ; : ; 1)$
$E2 \rightarrow X\ (: ; : ; 2)$
$E3 \rightarrow X\ (: ; : ; 3)$

Step 10: Apply *XOR* ($\oplus$) operation on *I1, I2, I3* and *E1, E2, E3* respectively

$I1 \oplus E1 \rightarrow E\_R$
$I2 \oplus E2 \rightarrow E\_G$
$I3 \oplus E3 \rightarrow E\_B$

Step 11: Concatenate the above three components
Enc_Img = ($E\_R \| E\_G \| E\_B$)

Encrypted images are published in cloud to store and share with the users. Unauthorized access can be defeated in such environment; Users should have both Shift Key and Key Image to decrypt. Owner is authentic to validate the legitimate user's requests and share the key for decryption.

5 Experimental Results

- **Visual Testing**:

The original image of Lena is taken from USC-SIPI Image database and Medical image from UCI machine Learning Repository as shown in Fig. 8. The cipher images of the proposed algorithm in Fig. 10 shows us clear encryption in the context of view and also in security to the cipher images of the existing algorithm in Fig. 9.

- **Information Entropy Analysis:**

The information entropy $H(X)$ is a statistical measure of uncertainty in communication theory. It is defined as follows:

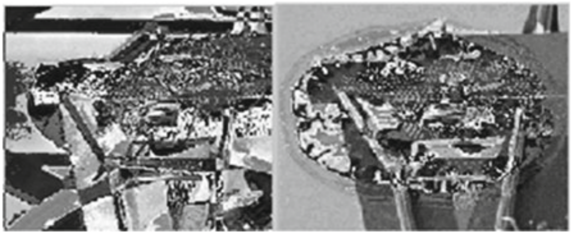


Fig. 9 Cipher images with existing algorithm

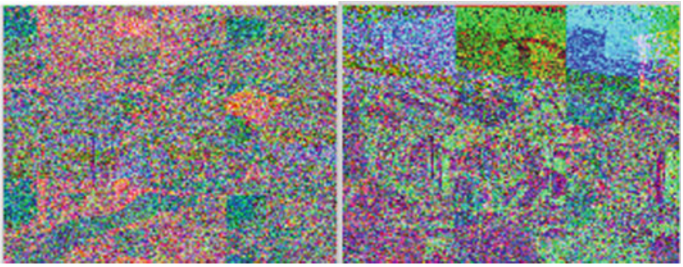


Fig. 10 Cipher images of Lena and medical with the proposed algorithm

Table 1 Entropy results of encrypted image

Entropy	Lena	Medical
Existing system	7.850	7.719
Proposed system	7.990	7.956

$$H(X) = -\sum_{i=0}^{255} p(x_i)\log2 p(x_i)$$

where X is a discrete random variable, $p(x_i)$ is the probability density function of the x_i.

Entropy for an encrypted image is measured because it is an important parameter to analyze encryption method. The entropy is an analytical parameter of ambiguity in transmission premise and it shows the extent of unpredictability in any transmission structure/scheme. We can get the perfect entropy $H(X) = 8$, corresponding to a truly random sample listed in Table 1.

Table 2 NPCR results of encrypted image

NPCR	Lena	Medical
Existing system	99.651	99.654
Proposed system	99.752	99.753

- **NPCR (Number of Pixel Change Rate)**

NPCR is a widely used security analysis in the image encryption for differential attacks. It calculates the sum of pixels changed for a cipher image during one pixel of original image changed, this value going to present in between 0 and 1. Higher value of NPCR is desirable for an economical ciphering of image as listed in Table 2.

E_1 and E_2 are the two encrypted images before and after the change of one pixel in original image appropriately

$$\text{Bipolar array}: B(m,n) = 0, \text{ if } E_1(m,n) = E_2(m,n)$$
$$1, \text{ if } E_1(m,n) \neq E_2(m,n)$$

$$NPCR = \sum B(m,n) \times 100\%$$
$$m,n - - - - - - - - - - - - - - - - -$$
$$T$$

6 Conclusions

Proposed hybrid encryption scheme uses SCAN pattern and key-based partitioning technique, which provides high security level. This scheme presents an efficient pixel equivalent transfiguration method with minimum loss and effectiveness with neighboring pixel resolution approach. Thus, maximum support against many cipher attack initiatives.

References

1. Pareek, N.K., Patidar, V., Sud, K.K.: Image encryption using chaotic logistic map. J. Image and Vision Comput. 926–934 (2006)
2. Zhou, Y., Cao, W., Philip Chen, C.L.: Image encryption using binary bitplane. J. Signal Process. **100**, 197–207 (2014)
3. Zhu, Z.L., Zhang, W., Wong, K. W., Yu, H.: A chaos-based symmetric image encryptionscheme using a bit-level permutation. J. Inf. Sci. **181**, 1171–1186, Elsevier (2010)
4. Zhang, C.Y., Zhang, W.X., Weng, S.W.: Comparison of two kinds of image scrambling methods based on LSB steganalysis. J. Inf. Hiding Multimedia Signal Process. **6**(4) (2015)
5. Yuan, H., Jiang, L.: Image Scrambling based on spiral filling of bits. J. Int. J. Signal Process., Image Process. Pattern Recognit. **8**(3), 225–234 (2015)

6. Bourbakis, N.: A language for sequential access of two-dimensional array elements. In: IEEE Workshop on LFA, pp. 52–58, Singapore (1986)
7. Rad, R.M., Attar, A., Atani, R. E.: A new fast and simple image encryption algorithm using scan patterns and XOR. J. Int. J. Signal Process. Image Process. Pattern Recognit. **6**(5), 275–290 (2013)
8. Li, C., Lo, K.T.: Cryptanalysis of an Image encryption scheme using cellular automata substitution and SCAN. In: PCM 2010, Springer-Verlag Berlin Heidelberg, LNCS 6297, pp. 601–610 (2010)
9. Panduranga, H.T., Naveen Kumar, S.K.: Hybrid approach for image encryption using SCAN patterns and carrier images. J. Int. J. Comput. Sci. Eng., **02**(02), 297–300 (2010)
10. Bourbakis, N., Alexopoulos, C.: A fractal based image processing language—formal modeling. J. Pattern Recognit. J. **32**(2), 317–338 (1999)
11. Alexopoulos, C., Bourbakis, N., Ioannou, N.: Image encryption method using a class of fractals. J. J. Electron. Imaging, pp. 251–259 (1995)
12. Wu, Y., Zhou, Y. Noonan, J.P., Agaian, S.: Design of image cipher using latin squares. Inf. Sci. **264**, 317–339 (2014)
13. Guosheng, Gu, Ling, Jie: A fast image encryption method by using chaotic 3D cat maps. Optik **125**, 4700–4705 (2014)
14. Fridrich, J.: Symmetric ciphers based on two-dimensional chaotic maps. Int. J. Bifurc. Chaos **8**(6), 1259–1284 (1998)
15. Saikumar, N., Bala Krishnan, R., Meganathan, S., Raajan, N.R.: An Encryption Approach for Security Enhancement in Images using Key Based Partitioning Technique. In: IEEE Conference on Circuit, Power and Computing Technologies (March, 2016)

An Unbiased Privacy Sustaining Approach Based on SGO for Distortion of Data Sets to Shield the Sensitive Patterns in Trading Alliances

B. Janakiramaiah, G. Kalyani, Suresh Chittineni and B. Narendra Kumar Rao

Abstract Distribution of data in the organizations which are having cooperative business is a common scenario for getting the benefits in the business. Modern technology in data mining has permitted to extract the unknown patterns from the repositories of enormous data. On the other hand, it raises problem of revealing the confidential patterns when the data is shared to the others. Privacy-preserving data mining is an emerging area for the research in the domain of security to deal with the need privacy for concerns of confidential patterns. The original database is to be transformed to conceal the confidential patterns. Along with concealing the confidential patterns, another important parameter that is to be addressed is attaining the balance between privacy and utility of the database which are generally inversely proportional to each other. Another challenging aspect in the transformation process is reducing the side effects, miss cost, and false rules that may occur by mining the transformed database. In this paper, a new method has been projected for concealing of association rules that are sensitive by carefully selecting the transactions for transformation using computational intelligence technique social group optimization. The

B. Janakiramaiah (✉)
Department of CSE, PVP Siddhartha Institute of Technology,
Vijayawada, Andhra Pradesh, India
e-mail: bjanakiramaiah@gmail.com

G. Kalyani
Department of CSE, DVR & Dr HS MIC College of Technology,
Vijayawada, Andhra Pradesh, India
e-mail: kalyanichandrak@gmail.com

S. Chittineni
Department of IT, Anil Neerukonda Institute of Technology & Sciences,
Vishakapatnam, Andhra Pradesh, India
e-mail: sureshchittineni@gmail.com

B. Narendra Kumar Rao
Department of CSCE, Sree Vidyanikethan Engineering College,
Tirupati, Andhra Pradesh, India
e-mail: narendrakumarrao@yahoo.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_17

outcome of the proposed approach is measured against the existing techniques based on computational intelligence methods to demonstrate the comparison of side effects with the proposed method.

1 Introduction

Data mining techniques are utilized for finding the knowledge in enormous volumes of information. In the prior period, individuals confronted the circumstances where data mining techniques reveal the sensitive knowledge also from the database which was not planned to be unveiled to others.

The issue of protecting the association rules is examined in [1, 5, 8, 16]. The preferred privacy is achieved with distortion-ased techniques in several ways in view of not to reveal any sensitive knowledge. Some of the usual methods for data distortion, which is illustrated in [7, 8], is the exchanging of the items among the transactions. Removal of a number of items in the dataset is alternative method illustrated in [3]. Blocking methods are examined in [1, 17, 18]. In blocking methods, the original values are substituted with a particular symbol in the proper transactions. Specially, in [1] a number of approaches are projected, which blocks in a different way, to attain the best probable results. In this case an opponent is thwarted from deducing a unknown item value in a particular transaction of the dataset, and in [6, 7] bayesian methods are used to avoid the implication of the unknown value by the opponent.

This paper concentrated on the issue of changing over a database into another one that masks some sensitive rules and at correspondingly safeguarding the general non-sensitive rules and patterns from the sanitized database. Knowledge sanitization is a strategy for making the sensitive knowledge secured from more extensive disclosures [9, 10]. The way towards changing over the original database into a transformed one for securing the sensitive knowledge is called sanitization [2].

2 Literature Review

Data distortion approaches operate by choosing a particular item for adding (or deleting) in some of the chosen transactions of the original database to conceal the association rules which are not to be revealed to others. Most commonly used approaches for data distortion comprises an alteration of values among the transactions [3, 7], and deletion specific items in the transactions of the dataset [14].

Lin et al. projected the sGA2DT, pGA2DT [12], and cpGA2DT [13] algorithms for concealing of item sets that is to be kept private by eliminating the transactions, based on genetic procedures. Each gene represents a solution comprising of transactions to be removed in the process of concealing.

Kalyani et al. [11] proposed a method for protecting the sensitive rules. The algorithm starts by selecting a rule of sensitive knowledge as the victim rule and an item of that rule with minimum influence on nonsensitive rules as the victim item.

Sensitive rules are clustered together by considering the measure of simlarity. The next step of the selecting the sensitive transactions is applied, to select the transactions based on sensitive items.

Bonam et al. [4] proposed a method for not disclosing the sensitive rules using an optimization technique particle swarm intelligence. The first transactions are chosen with PSO. The itemsets are taken from which the confidential rules are framed. Support of the confidential itemsets was altered as lower than the MST. The item which has occurrence value more is chosen and is eliminated from the chosen transactions to lower the support.

Chun-Wei Lin et al. [19] proposed an ant colony system (ACS)-based framework called ACS2DT, to diminish side effects and upgrade the usefulness of the transformed data set. The path of each ant in the population indicates a transaction of the database which is best suitable for deletion in the process of sanitization. The ACS comprises of the phases state transaction rule, pheromone updating rules, and selection process. To specify the ending of their tour, some termination conditions are defined by the authors. To estimate the effectiveness of the ant tours, a fitness function was defined as the weighted sum of side effects raised from the transformed database. To guide the ants in the best direction, a heuristic function was defined by the authors.

The authors of algorithms ACS2DT and RSIF-PSOW have evaluated their algorithms with other methods available in the literature of PPARM. Thus, the proposed method of this paper was compared with the methods, ACS2DT and RSIF-PSOW algorithms which are projected with an intention of concealing of sensitive itemsets or rules by affecting the nonsensitive itemsets or rules in a lesser extent. Hence, the proposed algorithm was evaluated by comparing with two existing methods.

3 Proposed Methodology

3.1 Problem Formulation

The problem statement of concealing sensitive association rules in association rule mining can be framed as, given a data set $\mathcal{D}$ which is to be shared to others, a selected set of rules identified from the result of association rule mining on $\mathcal{D}$ and a set of sensitive rules to be protected(SAR), create new data set $\overline{\mathcal{D}}$, with a constraints that the sensitive rules are not to be disclosed by mining $\overline{\mathcal{D}}$ and the nonsensitive rules of $\mathcal{D}$ must be disclosed by mining $\overline{\mathcal{D}}$ as like they are disclosed by mining $\mathcal{D}$. In this case, $\overline{\mathcal{D}}$ will be released to others for analyzing the patterns.

3.2 Algorithm for Concealing of Association Rules

The set of association rules considered as confidential are named as Rs. Consider the sensitive rule that is selected as Ri in the form $A \Rightarrow B$. The technique intentions at concealing $A \Rightarrow B$ by removing an item in A or B in the chosen transactions until support($A \Rightarrow B$) $<$ MST or confidence($A \Rightarrow B$) $<$ MCT. To moderate the side effects, the projected algorithm applies six steps. The steps were repeated until Rs is empty. The projected algorithm is shown in Algorithm 1.

In Step 1 of Algorithm 1 is meant for calculating the amount of transactions required for alteration for every of Rs and for each item of the rule estimating the value of impact nonsensitive rules. In Step 2, first chose the items with minimum value of impact as least impact items in every rule. Then identify the frequency of each item in least impact items. Among the items of least impact, the item with maximum frequency is selected as High_freq_item and then the a rule in Rs which is having this High_freq_item as one of the item is chosen for concealing process. Finally, the net value evaluated in Step 1 is considered as the number of transactions for modification to conceal this rule. In Step 3, the algorithm selects the other rules of Rs which are having similarity with the chosen rule of Step 2. In Step 4, based on the rules chosen in Step 3 subsets of the rule Ids are generated from the complete set to empty set. These subsets are stored which are used in the next step. In Step 5, subsets generated and stored in Step 4 are considered. In the rules of every subset, the items of the rules will be considered to decide the set of items which are sensitive used in the process of choosing the transactions for alteration as Item_list1. Then by taking database $\mathcal{D}$ and Item_list1, Social Group Optimization(SGO) is applied to select the transactions for alteration. The process of SGO is shown in Algorithm 2. The result of SGO is a set of transaction suitable best for alteration. Then in Step 6, the selected High_freq_item is removed from the returned transactions of SGO. After the alteration the set of rules R and the set of confidential rules Rs both will be updated based on the updated database.

SGO is an optimization technique designed by considering the social behavior and emotions of the persons in the society [15]. The process of SGO is depicted in Algorithm 2 consists of primarily two phases. In improving phase, Highly fitted member (Gbest) in the societal grouping attempts to spread the knowledge among the persons, which in turn, facilitate others to enhance their knowledge in the group. Each person receives knowledge from the groups best (Gbest) person. In acquiring phase, a member of the societal grouping intermingles with the finest member (Gbest) of the group and also intermingles with other members of the group for obtaining the knowledge. A member of the societal grouping attains new knowledge if the other member has higher knowledge when compared to him or her. The highest knowledgeable member will influence the others to acquire knowledge from him/her. Member of the societal grouping also learns something new from other members if they possess high knowledge than him or her in the societal grouping. The equations used in the improving and acquiring phase are shown in the Algorithm 2. After the improving and acquiring phase, the population will be updated based on the fitness function values of the population before and after the two phases. The fitness function used in the SGO is shown in Algorithm 2. The fitness function has designed based on the parameters sensitive item frequency and maximum decreased values of the items in the confidential rules. The best population is taken as the result if improving and acquiring phases. The process of improving and terminating phases are repeated until the specified iterations which is here taken as 100. In the improving phase a constant *c* will be used which is initialized to the value 2 in the implementation of the proposed algorithm. Acquiring phase make use of two random numbers generated using random number generation process during the implementation.

Algorithm 1: UHSSP(Unbiased Heuristic to Shield the Sensitive Patterns)

```
Data: Database $\mathcal{D}$ , Association rules R, MST, MCT, Rs
Result: A Sanitized Database $\overline{\mathcal{D}}$ .
begin
    repeat
        Step 1:
        for each rule $R_i \in Rs$ do
            Sitems = $(A \cup B)$ ;
            nt = Support(Sitems)-MST+1 ;
                /* no.of transactions required for hiding $A \Rightarrow B$ */
            Item_Impact_Values ($R_i$);
        Step 2:
        for each rule $R_i \in Rs$ do
            $Least\_Impact\_items = Least\_Impact\_items \cup Select\_min\_item(R_i)$;
        Calculate the frequencies of items in Least_Impact_items;
        High_freq_item = Select the item with highest frequency in Least_Impact_items;
        rule =$\{Ri/Ri \in Rs,\ Z \in Ri\ \&\ impactfactor(High\_freq\_item)\ is\ minimum\}$;
        Number_trans_modify= nt ; Items-List= $A \cup B$ ;
        Step 3:
        for every rule $R_i \in Rs$ do
            If (High_freq_item ∈ Ri)
            Rule_ID [pos]= i ;
            pos++;
        Step 4:
        for iter = 1 to $2^{pos}$ do
            Generate a subset s of Rule_ID[ ] values;
            Subset_index++;
            Sub_Sets[Subset_index]=w;
        Step 5:
        for iter = Subset_index to 1 step -1 do
            for each value ∈ Sub_Sets [c] do
                Items_List1 = items_List ∪ Rvalue;
            Selected_Transactions = SGO($\mathcal{D}$ , Item_List1);
        Step 6:
        Select the first no.of_trans_modify transactions and remove High_freq_item from
        those;
        Update(R);
        Update (Rs);
    until (Rs = Null);
```

Algorithm 2: Social Group Optimization(SGO)

```
Data: Database D, Sensitive items
Result: Set of sensitive transaction
begin
    Database D is divided into a K number of disjoint subsets of the database.
    Randomly select L number of subset databases
    Let $T_i, i = \{1, 2, 3, 4, .., n\}$ be the transactions of one subset database from L , this subset
    contains n of transaction and each transaction $T_j, j = \{A_1, A_2, A_3, A_m\}$, $T_j$ transaction
    contains m number of items.
    Improving phase:
    call Fitness Function(D,Rs);
    for i=1:n do
        for j=1:m do
            $T_{newij} = c * T_{oldij} + r * (gbest(j) - T_{oldij})$
    Acquiring phase:
    $gbest = max\{f(x_i), i = 1, 2, n\}$
    for i=1:n do
        Select one transaction randomly $T_r$;
        if $f(T_i) < f(T_r)$ then
            for j=1:m do
                $T_{newi,j} = T_{oldij} + r_1 * (T_{ij} - T_{r,j}) + r_2 * (gbest_j - T_{i,j})$
        else
            for j=1:m do
                $T_{newi} = T_{oldi,:} + r_1 * (T_{r,:} - T_{i,:}) + r_2 * (gbest_j - T_{ij})$
        Accept $T_{new}$ if it gives a better fitness function
```

Algorithm 3: Fitness Function

```
Data: Database D, Transactions
Result: fitness function values of transaction
begin
    $DC = C_i - MST * n + 1$
    $MDC = Max_{j=1}^{m} DC_{kj}$
    $FF = \sum_{k=1}^{p} \log \frac{n}{MDC - Support}$
return FF
```

4 Performance Metrics

The outcome of the proposed methodology is assessed by considering the following parameters.

A. ***Need To be Hidden (NTH)***:

This parameter specifies the amount of confidential rules revealed from the transformed database that are meant for not revealed from the transformed database. It is to be calculated in terms of percentage as

$$NTH = \frac{CR(\overline{\mathcal{D}})}{CR(\mathcal{D})} \tag{1}$$

In the above formula, CR($\overline{\mathcal{D}}$) stands for amount of confidential rules disclosed from the transformed database and CR($\mathcal{D}$) stands for amount of confidential rules disclosed from the original one.

B. ***Need To be Reveal-Missed (NTRM)***:
This parameter represents the amount of nonconfidential rules not disclosed from the transformed database which is originally meant for disclosure from the transformed database. It is to be calculated in terms of percentage as

$$NTRM = \frac{NCR(\overline{\mathcal{D}})}{NCR(\mathcal{D})} \tag{2}$$

In the above formula NCR($\overline{\mathcal{D}}$) stands for amount of nonconfidential rules disclosed from the transformed database and NCR($\mathcal{D}$) stands for amount of nonconfidential rules disclosed from the original one.

C. ***Not To be Generated (NTG)***:
This parameter specifies the amount of rules revealed from the transformed database which is not revealed from the original one. It is to be calculated in terms of percentage as

$$NTG = \frac{NR(\overline{\mathcal{D}})}{R(\mathcal{D})} \tag{3}$$

In the above formula NR($\overline{\mathcal{D}}$) stands for amount of new rules disclosed from the transformed database and R($\mathcal{D}$) stands for amount of rules disclosed from the original one.

D. ***Variation*** ($\mathcal{D}$, $\overline{\mathcal{D}}$):
This parameter specifies the variation between the original ad transformed databases in terms of percentage of alterations done in the process of concealing of confidential rules. It is to be calculated as

$$Variation(\mathcal{D}, \overline{\mathcal{D}}) = \sum_{k=1}^{CR} No.\ of\ \ alterations\ [k] \tag{4}$$

In the above formula CR stands for number of confidential rules selected for concealing and No.of alterations [k] stands for modifications done to conceal the kth confidential rule.

5 Experimental Evaluation

The performance of the proposed algorithm is assessed by selecting three real datasets Retailer, BMS-Web View-1, and Mushroom which are available in FIMI data repository. The evaluation is done in two ways. In the first case by varying the % MST & MCT values and in the second case by varying the % of SAR, the parameters NTRM, and variation($\mathcal{D}$, $\overline{\mathcal{D}}$) are estimated. The parameter NTH is 100% for all the algorithms, i.e., all the algorithms conceal the confidential rules completely.

Figures 1 and 2 shows the comparison results of variation($\mathcal{D}$, $\overline{\mathcal{D}}$) and NTRM by varying the % of MST & MCT with BMS-Web View-1 dataset. Figures 3 and 4 shows the comparison results of variation($\mathcal{D}$, $\overline{\mathcal{D}}$) and NTRM by varying the % of SAR with BMS-Web View-1 dataset.

Figures 5 and 6 shows the comparison results of variation($\mathcal{D}$, $\overline{\mathcal{D}}$) and NTRM by varying the % of MST & MCT with Mushroom dataset. Figures 7 and 8 shows the comparison results of variation($\mathcal{D}$, $\overline{\mathcal{D}}$) and NTRM by varying the % of SAR with Mushroom dataset.

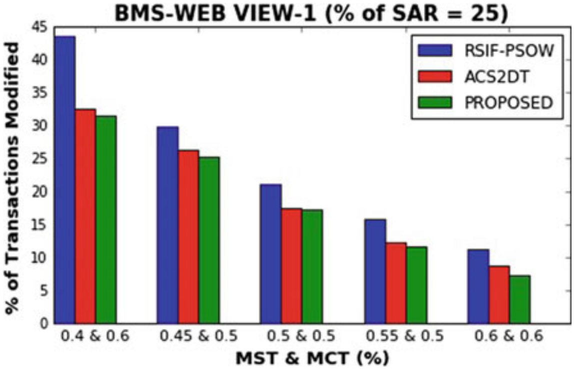


Fig. 1 With BMS-web view-1 dataset comparison of variation ($\mathcal{D}$, $\overline{\mathcal{D}}$) by varying % of MST & MCT

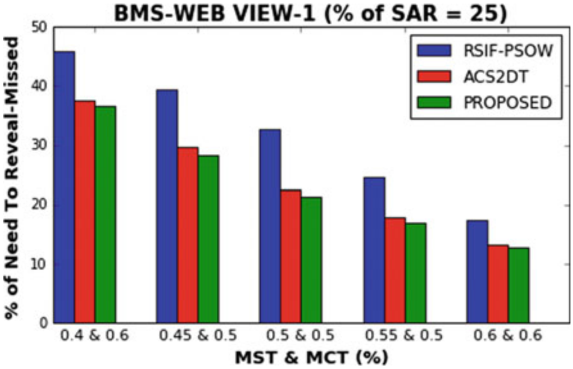


Fig. 2 With BMS-web view-1 dataset comparison of need to be reveal-missed by varying % of MST & MCT

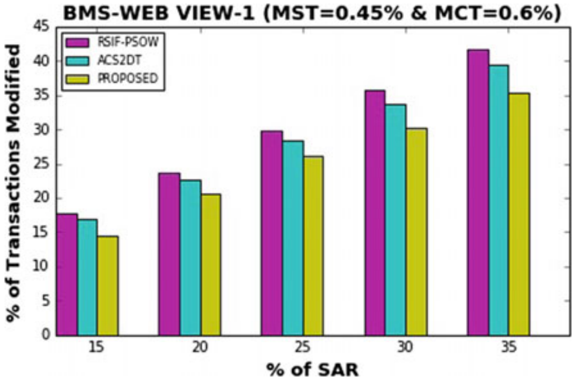


Fig. 3 With BMS-web view-1 dataset comparison of variation($\mathcal{D}$, $\overline{\mathcal{D}}$) by varying % of SAR

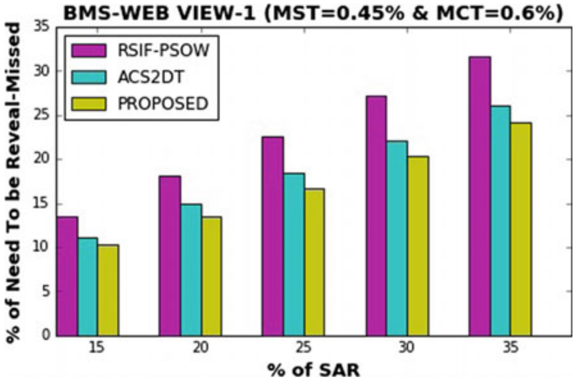


Fig. 4 With BMS-web view-1 dataset comparison of need to be reveal-missed by varying % of SAR

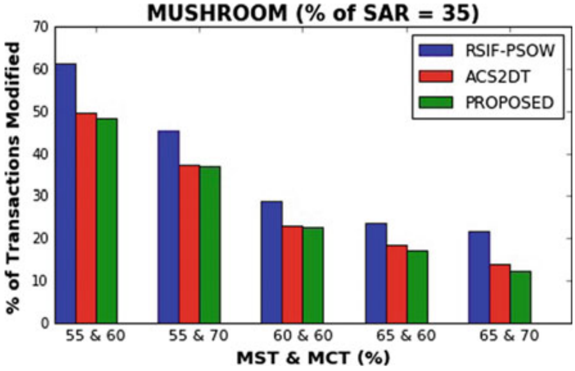


Fig. 5 With mushroom dataset comparison of variation ($\mathcal{D}$, $\overline{\mathcal{D}}$) by varying % of MST & MCT

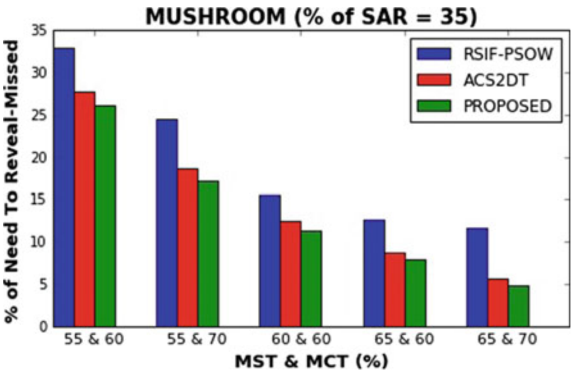


Fig. 6 With mushroom dataset comparison of need to be reveal-missed by varying % of MST & MCT

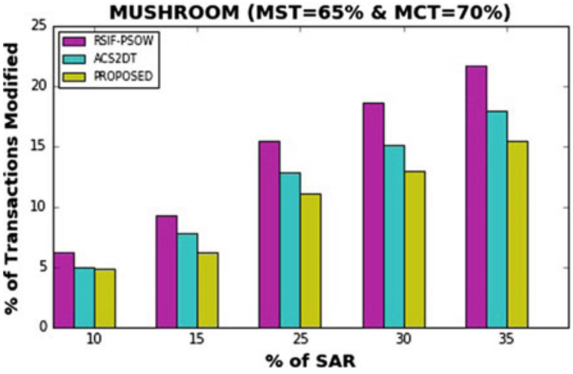


Fig. 7 With mushroom dataset comparison of variation $(\mathcal{D}, \overline{\mathcal{D}})$ by varying % of SAR

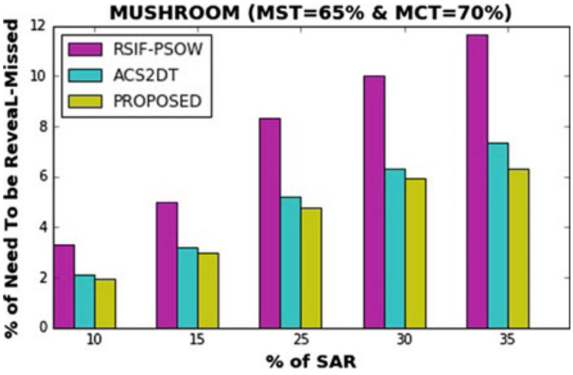


Fig. 8 With mushroom dataset comparison of need to be reveal-missed by varying % of SAR

Fig. 9 With mushroom dataset comparison of variation($\mathcal{D}$, $\overline{\mathcal{D}}$) by varying % of MST & MCT

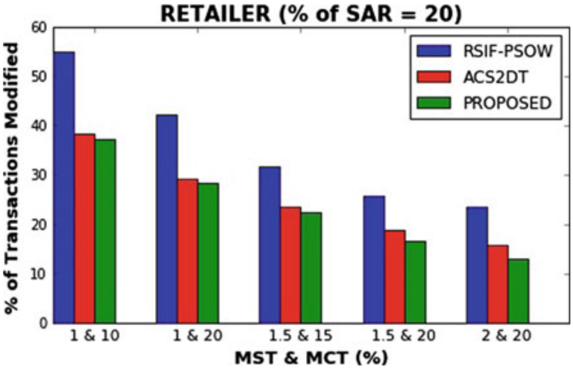


Fig. 10 With BMS-Web view-1 dataset comparison of need to be reveal-missed by varying % of MST & MCT

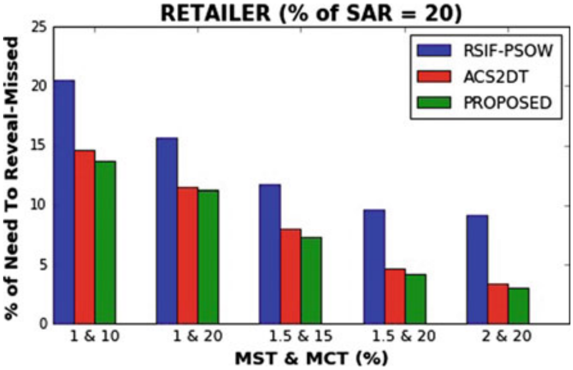


Fig. 11 With mushroom dataset comparison of variation ($\mathcal{D}$, $\overline{\mathcal{D}}$) by varying % of SAR

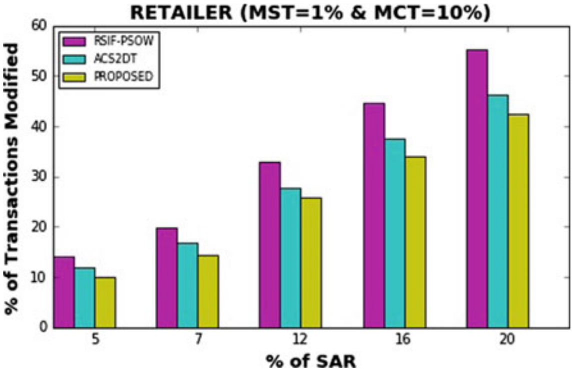
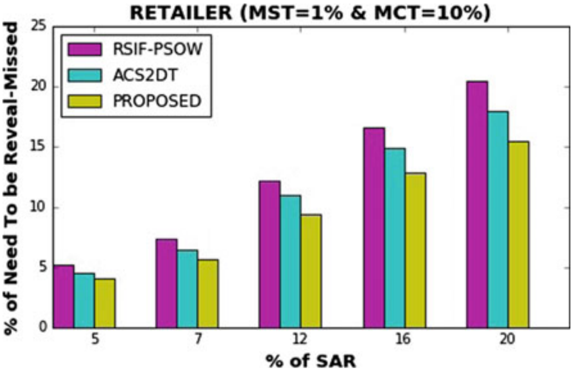


Fig. 12 With BMS-web view-1 dataset comparison of need to be reveal-missed by varying % of SAR



Figures 9 and 10 shows the comparison results of variation($\mathcal{D}$, $\overline{\mathcal{D}}$) and NTRM by varying the % of MST & MCT with Retailer dataset. Figures 11 and 12 shows the comparison results of variation($\mathcal{D}$, $\overline{\mathcal{D}}$) and NTRM by varying the % of SAR with Retailer dataset.

The experimental results indicated in the graphs significantly indicates the effectiveness of the proposed algorithm, i.e., reducing the NTRM(minimizing the loss of nonsensitive rules) and improving the utility of the dataset by minimizing the alterations in the data set. The proposed algorithm competes with the existing algorithms in the literature.

6 Conclusion

Privacy preserving data mining is an important issue in collaborated business applications. The expectation of privacy-preserving data mining will be hiding exactly confidential patterns with the goal that they cannot presented through with whatever standard data mining functionality. The idea of effective parameter is calculated in the suggested methodology might have been used to diminish the amount for alterations in the transactions and also will keep the characteristics of the original database. The methodology diminishes the amount of adjustments in the data by concealing more rules at a time to keep the usefulness of the given database. Test outcomes were studied by considering the on hand algorithms with respect to the measures NTH, NTRM, NTG, and variation ($\mathcal{D}$, $\overline{\mathcal{D}}$). The outcomes indicate that execution of the suggested algorithm may compete or equal the different existing methodologies based on computational intelligence.

References

1. Amiri, A.: Dare to share: protecting sensitive knowledge with data sanitization. Decis. Support Syst. **43**(1), 181–191 (2007)
2. Askari, M., Safavi-Naini, R., Barker, K.: An information theoretic privacy and utility measure for data sanitization mechanisms. In: *Proceedings of the second ACM conference on Data and Application Security and Privacy*, pp. 283–294. ACM (2012)
3. Atallah, M., Bertino, E., Elmagarmid, A., Ibrahim, M., Verykios, V.: Disclosure limitation of sensitive rules. In: *Knowledge and Data Engineering Exchange, 1999.(KDEX'99) Proceedings. 1999 Workshop on*, pp. 45–52. IEEE (1999)
4. Bonam, J., Reddy, A.R. Kalyani, G.: Privacy preserving in association rule mining by data distortion using pso. In: *ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India-Vol II*, pp. 551–558. Springer (2014)
5. Bonam, J., Reddy, R.: Balanced approach for hiding sensitive association rules in data sharing environment. Int. J. Inf. Sec. Priv. **8**(3), 39–62 (2014)
6. Chang L., Moskowitz, I.S.: Parsimonious downgrading and decision trees applied to the inference problem. In: *Proceedings of the 1998 workshop on New security paradigms*, pp. 82–89. ACM (1998)
7. Dasseni, E., Verykios, V.S. Ahmed K Elmagarmid, and Elisa Bertino. Hiding association rules by using confidence and support. In *International Workshop on Information Hiding*, pp. 369–383. Springer (2001)
8. Evfimievski, A., Srikant, R., Agrawal, R., Gehrke, J.: Privacy preserving mining of association rules. Inf. Syst. **29**(4), 343–364 (2004)
9. Kalyani, G., Chandra Sekhara Rao, M.V.P., Janakiramaiah, B.: Privacy-preserving classification rule mining for balancing data utility and knowledge privacy using adapted binary firefly algorithm. Arabian J. Sci. Eng. (2017)
10. Kalyani, G., Chandra Sekhara Rao, M.V.P., Janakiramaiah, B.: Decision tree based data reconstruction for privacy preserving classification rule mining. Informatica **41**(3) (2017)
11. Kalyani, G., Chandra Sekhara Rao, M.V.P., Janakiramaiah, B.: Particle swarm intelligence and impact factor-based privacy preserving association rule mining for balancing data utility and knowledge privacy. Arabian J. Sci. Eng. 1–18 (2017)
12. Lin, Chun-Wei, Hong, Tzung-Pei, Yang, Kuo-Tung, Wang, Shyue-Liang: The ga-based algorithms for optimizing hiding sensitive itemsets through transaction deletion. Appl. Intell. **42**(2), 210–230 (2015)
13. Lin, C.-W., Zhang, B., Yang, K.-T., Hong, T.-P.: Efficiently hiding sensitive itemsets with transaction deletion based on genetic algorithms. Sci. World J. (2014)
14. Oliveira, S.R.M., Zaïane, O.R.: Protecting sensitive knowledge by data sanitization. In: *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pp. 613–616. IEEE (2003)
15. Satapathy, Suresh, Naik, Anima: Social group optimization (sgo): a new population evolutionary optimization technique. Complex & Intell. Syst. **2**(3), 173–203 (2016)
16. Saygin, Yücel, Verykios, V.S., Clifton, C.: Using unknowns to prevent discovery of association rules. Acm. Sigmod Rec. **30**(4), 45–54 (2001)
17. Saygin, Y., Verykios, V.S., Elmagarmid, A.K.: Privacy preserving association rule mining. In *Research Issues in Data Engineering: Engineering E-Commerce/E-Business Systems, 2002. RIDE-2EC 2002. Proceedings. Twelfth International Workshop on*, pp. 151–158. IEEE (2002)
18. Verykios, V.S., Elmagarmid, A.K., Bertino, E., Saygin, Y., Dasseni, E.: Association rule hiding. IEEE Trans. Knowledge Data Eng. **16**(4), 434–447 (2004)
19. Wu, J.M.-T., Zhan, J., Lin, C.-W.: Ant colony system sanitization approach to hiding sensitive itemsets. IEEE Access (2017)

Contrast-Enhanced Recursive Visual Cryptography Scheme Based on Additional Basis Matrices

Thomas Monoth

Abstract Security and contrast are the two key parameters in visual cryptography schemes. Security condition is fulfilled if each share reveals no information of the secret image and the secret image cannot be obtained if there are k-1 shares together. The security and reliability of the visual cryptography scheme can be greatly enhanced with recursion. The contrast of the reconstructed secret image is less compared to the existing visual cryptography schemes. The proposed model provides improved contrast and reduces the noise in the recreated image without any additional computation. This scheme uses the new pixel patterns to enhance the contrast of the decrypted image. The contrast of the recursive visual cryptography scheme can be enhanced based on new pixel patterns for white pixels.

1 Introduction

Visual cryptography (VC) is a new type of cryptographic model proposed by Naor and Shamir [1]. In VC, decryption can be performed by human eyes without any computation. Visual cryptography scheme (VCS) divides the secret image into n shares from which any k shares can recreate the secret image. The main attraction of VCS is decryption that can be performed without any computations. Gnanaguruparan and Kak [2] invented the recursive hiding of secrets in visual cryptography scheme. This scheme allows extra secrets that can hide in a share and also reduces the network load [3–5].

T. Monoth (✉)
Department of Computer Science, Mary Matha Arts & Science College, Kannur University,
Mananthavady 670645, Kerala, India
e-mail: tmonoth@yahoo.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_18

2 Recursive Visual Cryptography Scheme

Security is one among the key factor of visual cryptography scheme. Security is satisfied if no information can be obtained from each share and cannot be decrypted from k-1 shares [6]. Here, we explained security-enhanced recursive visual cryptography scheme (RVCS). In RVCS, the image is encrypted into shares and sub-shares based on recursion. VCS with recursion, the reliability, and security can greatly be enhanced [7–10].

2.1 The Model

Let $P = \{p1, p2, \ldots, pn\}$ be n participants and 2^P denote the subsets of P. Let ΓQual ⊆ 2^P and ΓForb ⊆ 2^P such that ΓQual ∩ ΓForb = Ø. In the first phase of the encryption process, ΓQual are referred to as qualified sets and ΓForb as forbidden sets [11].

In the second phase of encryption, let $pi = \{pi1, pi2, \ldots\ldots pin\}$, for $i = 1$ to n. Then, the qualified sets ΓQuali ⊆ 2pi and forbidden sets ΓForbi ⊆ 2pi such that ΓQuali ∩ ΓForbi = Ø. This process can be repeated up to the desired security and contrast. The value of n can be different at each phase or depends on the number of shares/participants required at each phase. In the decryption process, secret image (SI) can be reconstructed as per Eq. (1).

In the first phase,

$$\mathrm{SI} = \sum_{i=1}^{k} pi \tag{1}$$

where k participants can reconstruct the secret image. The value of k is different for different VCSs. In this, each of the participants (for example, pi) is reconstructed using Eq. (2).

$$\mathrm{p_i} = \sum_{j=1}^{k} p_{ij} 1 \leq i \leq n \tag{2}$$

2.2 The Construction of Recursive Visual Cryptography Scheme

The recursive visual cryptography scheme is explained based on 2-out-of-2 VCS with two levels of encryptions. The image is encrypted into two shares at first level using 2-out-of-2 VCS [9, 10]. In the second level, each share is further encoded into

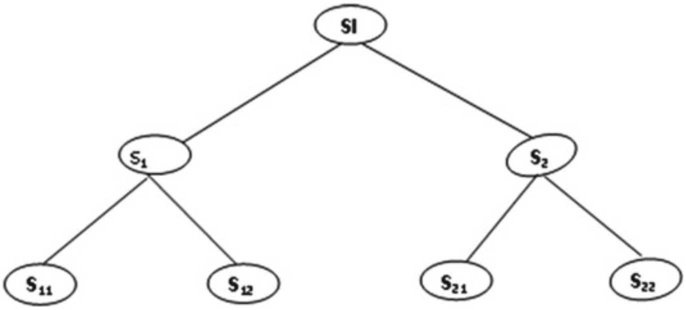


Fig. 1 Tree representation for *2-out-of-2* VCS with recursion using two levels of encryption

two shares using 2-out-of-2 VCS. This encryption process can be represented by a tree as shown in Fig. 1:

From Fig. 1, SI is encrypted into two shares S1 and S2. Then, from the first level the share S1 is again encrypted into two shares S11 and S12 and the share S2 into S21 and S22. In the decryption process, the SI is reconstructed by overlapping shares in different ways [10]. That is,

$$\begin{aligned} \text{SI} &= \text{S1} + \text{S2} \\ \text{SI} &= \text{S1} + \text{S21} + \text{S22} \\ \text{SI} &= \text{S2} + \text{S11} + \text{S12} \\ \text{SI} &= \text{S11} + \text{S12} + \text{S21} + \text{S22} \end{aligned}$$

Four different manners decrypt the image using visual cryptography scheme with recursion. But the existing visual cryptography scheme can recreate secret image in only one way. Therefore, visual cryptography scheme with recursion provides greater security and reliability than existing VCS. Analyze the security in VCS with recursion by comparing it with Naor and Shamir VCS. Using different levels of encryptions, the security and reliability have been greatly improved and the crypt-analysis is impossible [12].

2.3 Example

In RVCS, the experiments are based on 2-out-of-2 VCS with two levels of encryptions. Figure 2 shows RVCS [10]:

From Fig. 2, the reconstructed images are (b), (c), and (d) using OR (+) operations.

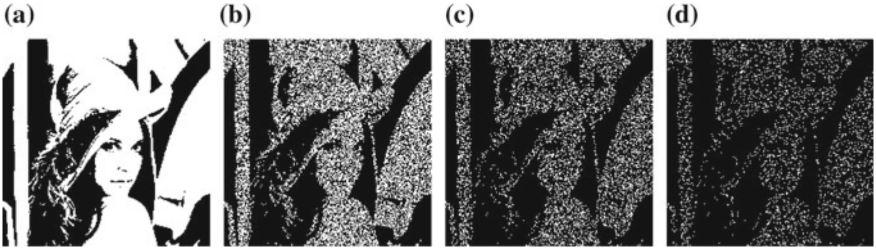


Fig. 2 The *2-out-of-2* RVCS for the image SI: **a** SI **b** $S_1 + S_2$ **c** $S_1 + S_{21} + S_{22}$ **d** $S_{11} + S_{12} + S_{21} + S_{22}$

3 Proposed Method

The proposed scheme RVCS with additional basis matrix (ABM) method enhances the contrast of RVCS [12]. The ABM is used for the generation of shares at each phase. Using ABM and recursion together, the security can be improved and contrast can be equivalent to Naor and Shamir VCS. The construction and implementation of ABM are explained below:

3.1 The Proposed Model

In RVCS, the ABM can be represented as AS^0 [13]. The basis matrix AS^0 is used to share white pixels in the share. The AS^0 is defined by *n x m* matrix, $AS^0 = [as_{ij}]$, where

$as_{ij} = 1 \Leftrightarrow$ the *j*th subpixel in the *i*th share is black.
$as_{ij} = 0 \Leftrightarrow$ the *j*th subpixel in the *i*th share is white.

Formula 1 (*Additional Relative Difference*) [14] Let $[s^0_{ij}]$ and $[as^0_{ij}]$ are two *n x m* matrices.

The contrast is calculated using Eq. (3).

$$\alpha * = (\alpha_1 + \alpha)/2 \tag{3}$$

where

$\alpha_1 = (\omega_H(S^1) - \omega_H(AS^{0)})/m$

$\alpha = (\omega_H(S^1) - \omega_H(S^{0)})/m$

ω_H *(S^1)* is the *Hamming weight* of S^1 and ω_H ($AS^{0)}$ is the Hamming weight of AS^0.

Formula 2 (*Additional Contrast*) Let $\alpha *$ and *m* are the additional relative difference and pixel expansion, respectively.

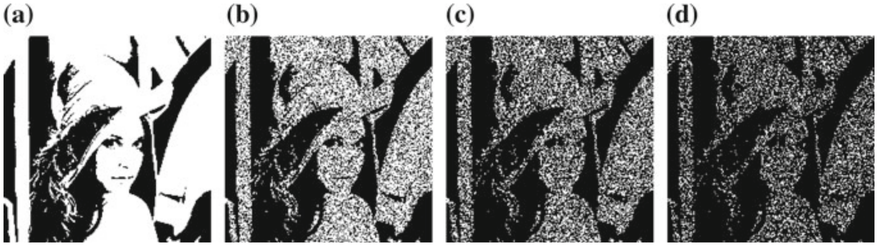


Fig. 3 The *2-out-of-2* RVCS: **a** SI, **b** $S_1 + S_2$ **c** $S_1 + S_{21} + S_{22}$ **d** $S_{11} + S_{12} + S_{21} + S_{22}$

Table 1 The details of the pixels in SI for the *2-out-of-2* RVCS

Secret image	Total columns	Total rows	Total black pixels	Total white pixels	Total pixels
SI	200	200	16,665	23,335	40,000
$S_1 + S_2$	200	200	28,273	11,727	40,000
$S_1 + S_{21} + S_{22}$	200	200	34,188	5812	40,000
$S_2 + S_{11} + S_{12}$	200	200	34,177	5823	40,000
$S_{11} + S_{12} + S_{21} + S_{22}$	200	200	37,162	2838	40,000

Therefore,

$$\beta^* = \alpha^*.m, \quad \beta^* \geq 1$$

3.2 Experimental Results

In order to demonstrate the RVCS with ABM, the experiments were explained with 2-out-of-2 VCS based on two-level encryption with ABM. Figure 3 shows RVCS with ABM applied to secret image.

3.3 Analysis of Experimental Results

Analyze the security in RVCS by comparing it with Naor and Shamir VCS. Using different levels of encryptions, the security and reliability have been greatly improved. Here, we using two levels of encryption, cryptanalysis becomes impossible.

The details of pixels in the secret images and the reconstructed images obtained by stacking the shares in different ways are shown in Table 1. The contrast of the RVCS is analyzed with the help of graphs (Fig. 4) [15].

Fig. 4 The graphical representation of pixel details of SI based on RVCS

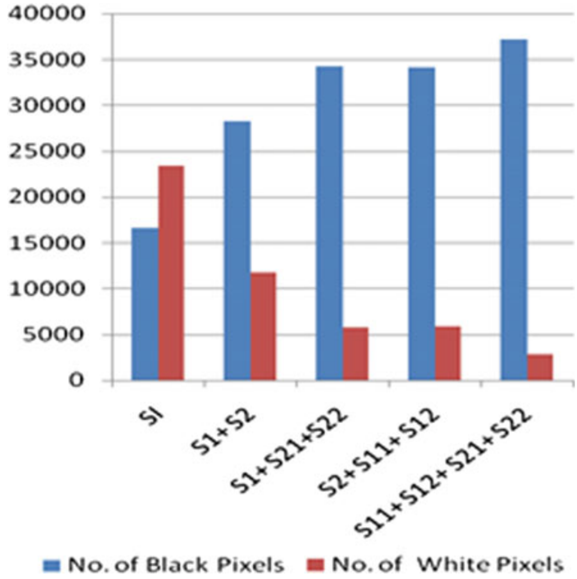


Table 2 The pixels in SI for 2-out-of-2 RVCS with ABM

Secret image	Total columns	Total rows	Total black pixels	Total white pixels	Total pixels
SI	200	200	16,665	23,335	40,000
$S_1 + S_2$	200	200	25,079	14,921	40,000
$S_1 + S_{21} + S_{22}$	200	200	30,396	9604	40,000
$S_2 + S_{11} + S_{12}$	200	200	30,513	9487	40,000
$S_{11} + S_{12} + S_{21} + S_{22}$	200	200	33,873	6127	40,000

From the graphs (Fig. 4), the number of white pixels is reduced considerably in reconstructed secret image by stacking three shares ($S_1 + S_{21} + S_{22}$ or $S_2 + S_{11} + S_{12}$) compared to stacking two shares ($S_1 + S_2$) in both images. Similarly, in the second level, also white pixels are reduced in the reconstructed image by stacking four shares ($S_{11} + S_{12} + S_{21} + S_{22}$) compared to stacking three shares ($S_1 + S_{21} + S_{22}$ or $S_2 + S_{11} + S_{12}$). When the number of shares stacked is increased, this will reduce the contrast. In order to minimize the contrast loss in recursive visual cryptography scheme, use ABM method. The number of the pixels in the images and the image can be obtained by stacking the shares in different ways based on RVCS with ABM as shown in Table 2.

By analyzing the results, we compared RVCS with ABM and RVCS using graphs.

From the graphs (Figs. 4 and 5), the white pixels are increased in the reconstructed images in RVCS with ABM in various ways compared with RVCS. Therefore, one

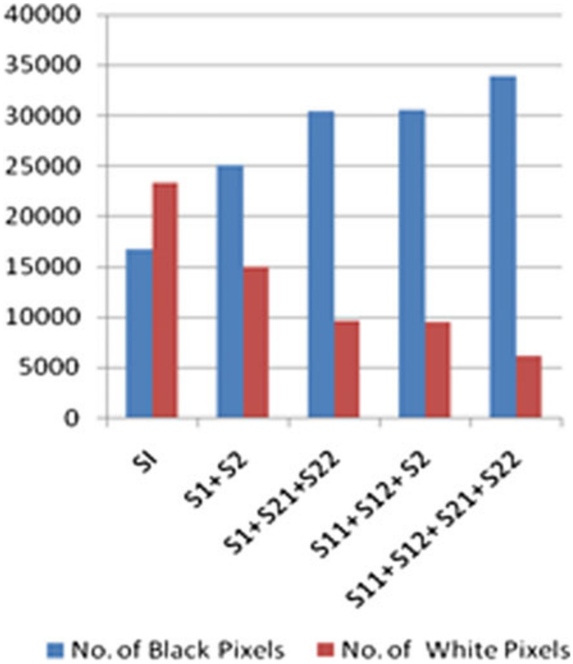


Fig. 5 The graphical representation of pixel details of SI in RVCS with ABM

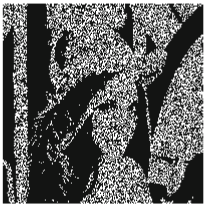
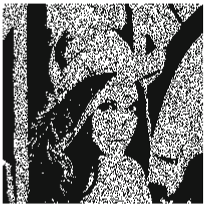
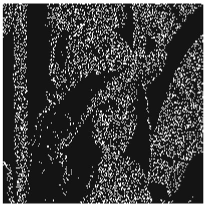
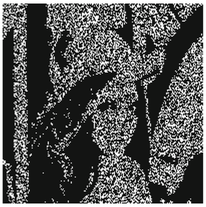
can see that RVCS with ABM achieves better contrast than RVCS. Next, analyze the reconstructed images in RVCS and RVCS with ABM.

By comparing the reconstructed images in Table 3, we see that RVCS with ABM achieves better and clearer image than RVCS. We proved that RVCS with ABM scheme obtains the same contrast and more security compared to Naor and Shamir VCS.

4 Conclusions

A new scheme is presented for RVCS with enhanced security and contrast with simple examples. In Naor and Shamir scheme, only one type and one level of VCS are used for the encoding and decoding of the SI, but in VCS with recursion different VCSs and more than one level of encoding and decoding can be used. Therefore, the security and reliability are enhanced. Contrast-enhanced recursive visual cryptography scheme with ABM is presented here. The RVCS with ABM provides almost the same contrast but better security compared to Naor and Shamir VCS. This approach can easily be extended to *k*-out-of-n VCS.

Table 3 The comparison of reconstructed images between RVCS and RVCS with ABM

Stacked shares	RVCS	RVCS with ABM
$S_1 + S_2$		
$S_1 + S_{21} + S_{22}$		

Note

This paper is a part of my Ph.D. thesis and the extension of my other papers. Some text in this paper is taken from my papers cited in the references [7–10, 12–14].

References

1. Naor, M., Shamir, A.: Visual cryptography, advances in cryptology-Eurocrypt'94. LNCS **950**, 1–12 (1995)
2. Gnanaguruparan, M., Kak, S.: Recursive hiding of secrets in visual cryptography. Cryptologia **26**(1), 68–76 (2002)
3. Parakh, A., Kak, S.: Recursive secret sharing for distributed storage and information hiding. In: Proceedings of the IEEE 3rd International Symposium on Advanced Networks and Telecommunication Systems (ANTS), pp. 1–3 (2009)
4. Parakh, A., Kak, S.: Space efficient secret sharing: a recursive approach, Cryptology ePrint Archive: Report 2009/365 (2009)
5. Parakh, A., Kak, S.: A tree based recursive information hiding scheme. In: Proceedings of the IEEE Communication and Information System Security Symposium, Cape Town, South Africa (2010)
6. Revenkar, P.S., Anjum, A., Gandhare, W.Z.: Survey of Visual Cryptography Schemes. Int. J. Security & Its Appl. **4**(2) (2010)
7. Monoth, T., Anto, B.P.: Recursive visual cryptography using random basis column pixel expansion. In: Proceedings of the IEEE International Conference on Information Technology (ICIT), Rourkela, Orissa, pp. 41–43 (2007) (IEEE Computer Society, ISBN : 0-7695-3068-0)
8. Monoth, T., Anto, B.P.: Tamperproof transmission of fingerprints using visual cryptography schemes. J. Procedia Comput. Sci. **2**, 143–148 (2010) (Elsevier, Netherlands, ISSN : 1877-0509)
9. Monoth, T., Anto, B.P.: Contrast-enhanced visual cryptography schemes based on additional pixel patterns. In: Proceedings of the IEEE International Conference on Cyber Worlds (CW

2010), NTU, Singapore, pp. 171–178 (2010) (IEEE Computer Society, ISBN : 978-0-7695-4215-7)

10. Monoth, T., Anto, B.P.: Security-enhanced visual cryptography schemes based on recursion. Commun. Comput. Inf. Sci. (CCIS) **140**, 255–262 (2011) (Springer-Verlag, Germany, ISSN : 1865-0929)
11. De Bonis, A., De Santis, A.: Randomness in visual cryptography. LNCS **1770**, 626–638 (2000)
12. Monoth, T., Anto, B.P.: Analysis and design of tamperproof and contrast-enhanced secret sharing based on visual cryptography schemes. Ph.D thesis, Kannur University, Kerala, India (2012) (http:// shodhganga.inflibnet.ac.in)
13. Monoth, T.: Anto, B.P.: Achieving Optimal Contrast in Visual Cryptography Schemes Without Pixel Expansion. Int. J. Recent Trends in Eng. (IJRTE), **1**(1), 468–471 (2009) (www.academypublisher.com, ISSN: 1787-9617)
14. Monoth, T., Anto, B.P.: Contrast-enhanced visual cryptography schemes based on perfect reconstruction of white pixels and additional basis matrix. Adv. Intell. Syst. Comput. (AISC): **412**:361–368 (2016). (Springer Science, Germany, ISSN : 2194-5357)
15. Yang, C.-N., Chen, T.-S.: Colored visual cryptography scheme based on additive color mixing. Pattern Recognit. **41**(10), 3114–3129 (2008)

Authenticity and Integrity Enhanced Active Digital Image Forensics Based on Visual Cryptography

T. E. Jisha and Thomas Monoth

Abstract Digital image forensics is an emerging research area which focuses on validating the authenticity, integrity of images, and detection of image forgeries. Digital image watermarking and digital signatures are presently used in active digital image forensics, but these methods provide only the authenticity and certain level of integrity and robustness. In this paper, we proposed a new method for active digital image forensics based on visual cryptographic schemes which provide improved authenticity, confidentiality, integrity, and robustness. Here, we have presented an overview of active digital image forensics and visual cryptographic schemes which have been discussed based on several studies. We have also discussed the demerits of the existing model and proposed a new model for image authentication and tampering detection. The proposed method is demonstrated with a simple example.

1 Introduction

Digital image forensics is the analysis of image authenticity and its content, forgery detection, and source identification. Forensic image analysis is the application of computer forensics to interpret the image and its content in judicial affairs. In our day-to-day life, images and videos are used to represent a common source of proof. Image in dailies is commonly accepted as an evidence of the reported news. Like this, image or video supervision records can comprise primary probationary material under judiciary [1].

In this era, social networks being a dominant communication tool, it is essential to design and execute methods that ensure the authenticity of the broadcast information.

T. E. Jisha (✉)
Department of Information Technology, Kannur University, Kannur 670 567, Kerala, India
e-mail: jishatevinoy@gmail.com

T. Monoth
Department of Computer Science, Mary Matha Arts & Science College, Kannur University, Mananthavady, Wayanad 670 645, Kerala, India
e-mail: tmonoth@yahoo.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_19

Since the current mobile devices have the facility to take large number of images easily, the images are treated as one of the most dominant communication media and shared documents at these social networks. In this perspective, it is of paramount importance to develop methods for verifying image authenticity [2]. Because of the limited success of the present investigation system to prosecute the criminals, the current investigation methods and systems need to be thoroughly understood and made stronger in order to control the cybercrime. Digital image forensics is under this category. The digital images in the cyberspace are powerful resources of information and are very easy to edit and transmit. The intruders can very easily modify the digital images with advanced image editing software. The authenticity and integrity of the digital images are of great concern to the society. Therefore, the study and research to ensure authenticity and integrity of digital images are incredible [3].

For ensuring the authenticity and integrity of digital images, different techniques are available today. The three main areas of digital image forensics are source identification (identify the capturing device), differentiate the images (whether it is natural or synthetic), and tampering detection. The methods used for authentication and integration of the image are subclassed into two: active and passive. In active method, digital signature is inserted into a digital image to determine the authenticity and tampering based on whether the retrieved signature is corrupted or not, whereas in passive digital image forensics, the authenticity and tampering are detected through feature extraction using unsupervised learning [4].

Digital image editing and transmission can be performed effortlessly with the usage of advanced and reliable image editing software, and the tampering detection is difficult by human eyes. With the vast usage of images in various fields such as journalism, crime detection, and judiciary, image forgery is a great hazard to the security of citizens. The image manipulation detection is a vital issue and therefore the implementation of trustworthy methods is necessary for image integrity investigation, source identification, and detection of image tampering [5].

Naor and Shamir [6] introduced a new kind of image sharing scheme, known as visual cryptography (VC). It has the capability of recovering the secret image with no computation. In k-out-of-n Visual Cryptography Scheme (VCS), secret image is encrypted into n shares from which any k or more shares can reconstruct the secret image. The main advantage of VCS is that reconstruction can be performed without any computations and the secret image can be revealed with OR operation. These characteristics make visual cryptography valuable especially for low computational devices [7, 8].

The foremost intention of the proposed model is to evaluate the authenticity, integrity, and tampering detection of the digital image. The coming sections of the paper are as follows. In the next sector, we have described a study on active digital image forensics model and its limitations. Section 3 demonstrates the proposed model with a simple example. Section 4 draws the conclusion.

2 The Active Digital Image Forensic Model

Uses of digital images as evidence for decision-making or judgments and as support for a scientific argument are examples where not only ownership of the images is required to be established, but it is equally important to establish their authenticity. Digital image watermarking and digital signatures have been used as active methods to restore the lost trust in digital images. These approaches are used in the digital image for the purpose of assessing the authenticity and integrity. Digital image watermarking belongs to the class of active approach for image forensics as it requires the knowledge of the authentication code and the method used to embed it into the image. The hidden information is generally imperceptible and robust against most of the intended and unintended attacks like histogram processing, compression, rotation, cropping, resampling, filtering, noise addition, etc. But a major disadvantage of active techniques is that they require manipulation of the original image either during capturing or during storage. Moreover, the need of generating the digital signature or watermark before saving the images calls for specially equipped image capturing devices. Thus, the use of digital signatures and watermarking as image forensic tools is not widely adopted. Based on the embedding domain, active approaches are further categorized into spatial and frequency domains. Figure 1 shows the generic active image forensic method [9, 10].

2.1 *Disadvantages of Active Digital Image Forensic Model*

The major disadvantages of active digital image forensic model using watermarking are that over-sized watermarks cover larger portions of an image and blur the image's contrast. Small watermarks can easily be removed using image editing software.

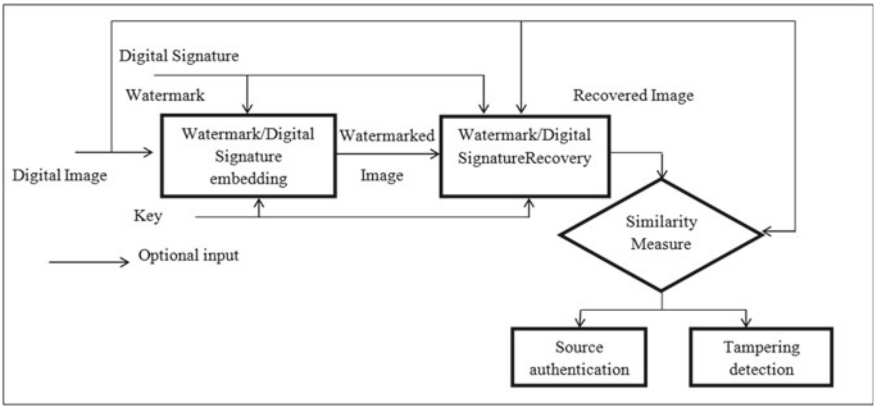


Fig. 1 Active forensic model

Embedding digital watermarks to digital image can be a time-consuming process [11].

The major drawbacks of active digital image forensics using digital signature are that to keep the private key safely is a difficult process. The time complexity is very high for key generation and verification of digital signature, and the storage of all previous keys is another issue. The digital signature does not provide the confidentiality of the image. Due to these demerits of the existing model, we need a new model for image authentication and tampering detection.

Digital signatures and watermarking are used in the existing model. It provides authenticity but not provides perfect confidentiality, reliability, integrity, and robustness. The computational complexity of digital image forensics is very high based on digital signatures and watermarking method.

3 The Proposed Active Digital Image Forensic Model

The proposed model based on VCS enhances the authenticity, confidentiality, integrity, and robustness of active digital image forensics model. In this method, the digital image is divided into *n* shares where the single share gives no information about the secret image. The *k* shares can be transmitted via secure channel and *n-k* shares can be transmitted via insecure channel. By stacking any *k* shares, the secret image can be reconstructed, and stacking *k-1* shares cannot reconstruct the original image. Using XOR operation for regeneration of the original image, we can avoid the contrast degradation. Visual cryptography scheme provides a secure way to send secret image with less computation [7, 12]. The schematic diagram for preparation and transmission of digital image based on visual cryptography is shown in Fig. 2.

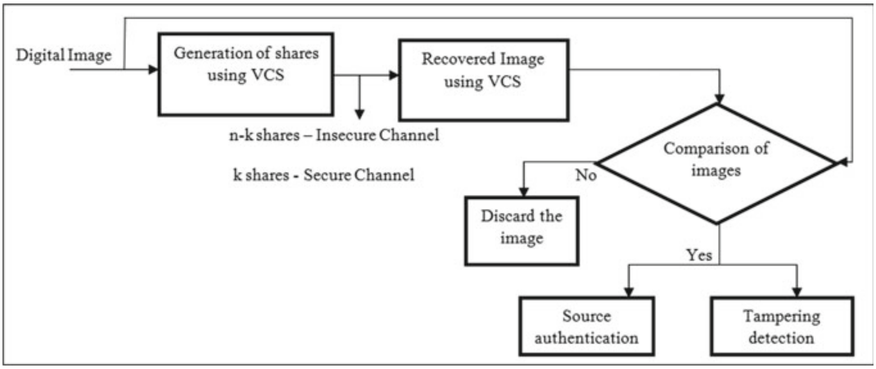

Digital Image
Generation of shares using VCS
Recovered Image using VCS
n-k shares – Insecure Channel
k shares - Secure Channel
Comparison of images
No
Discard the image
Yes
Source authentication
Tampering detection

Fig. 2 Proposed active forensic model

Figure 2 shows the proposed model of authenticity and integrity enhanced active digital image forensics. In this system, the image is encrypted into shares through VCS. The individual shares give no information about the image. This model can prevent any type of cryptographic attack. The major attractions of this model are as follows:

- The method provides perfect security and easy implementation.
- It has very low computational complexity.
- It provides perfect authenticity, confidentiality, integrity, and robustness.
- Decryption can be done without any key and knowledge of cryptography.
- It ensures high reliability.

3.1 Experimental Results

The experiments are based on *2-out-of-3* VCS. In this experiment, the images are considered as binary. In the encryption process, the image files are encrypted into three shares and the original secret image is revealed by overlapping any two shares using XOR operation. Figure 3 depicts the experimental results. The scheme provides perfect security, easy implementation, and very low computation decryption process.

Figure 3 shows the perfect reconstructed images (e–g) based on XOR ($\oplus$).

3.2 Analysis of Experimental Results

We analyzed the authenticity, integrity, computational complexity, computational cost, and robustness of the reconstructed digital image by comparing it with existing model based on digital watermarking [13–18] and digital signature [19–22]. Using visual cryptography scheme [7, 12, 23, 24] for both encryption and decryption of the image, the security and reliability have been greatly improved. The cryptanalysis becomes very complex or even impossible. We compared the existing active digital image forensics model and proposed model with respect to robustness, authenticity, integrity, security, computational complexity, and computational cost as shown in Table 1. On analyzing, it has been found that the visual cryptography scheme will play a pivotal role in active digital image forensics.

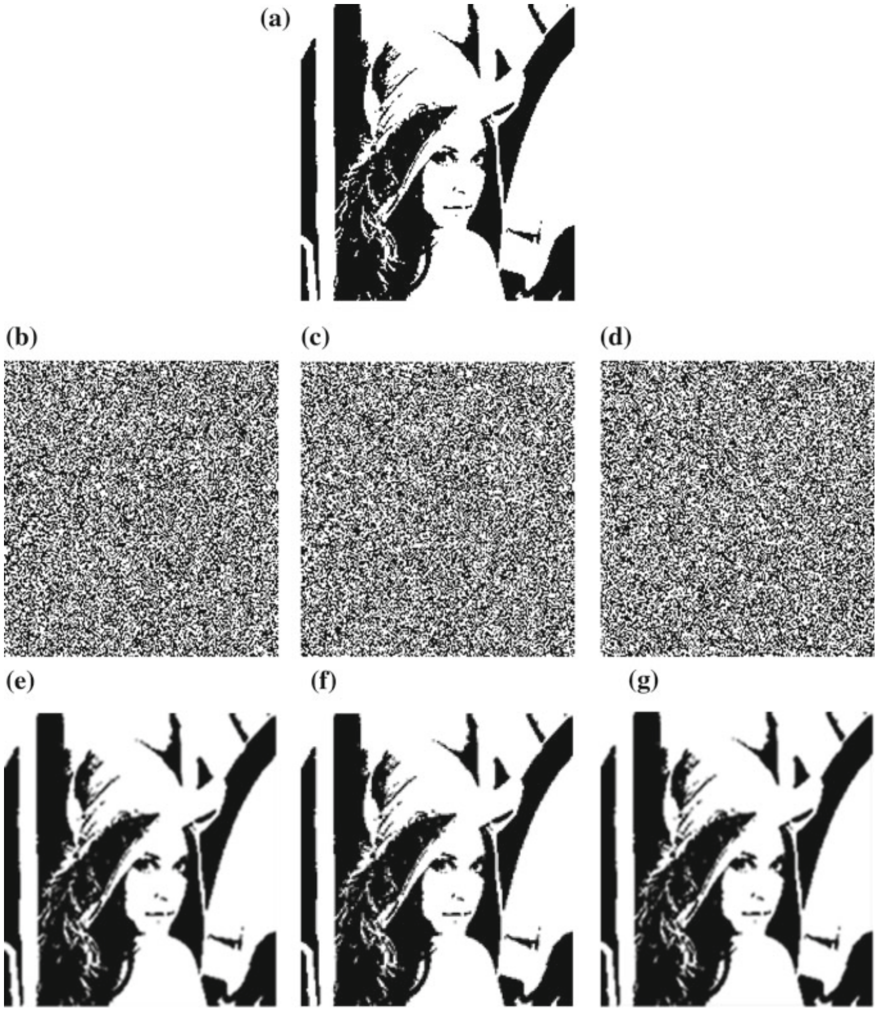


Fig. 3 **a** Secret image, **b** S_1, **c** S_2, **d** S_3, **e** $S_1 \oplus S_2$ (f) $S_1 \oplus S_3$ (g) $S_2 \oplus S_3$

4 Conclusions

In this paper, we presented a new method for active digital image forensics based on visual cryptography schemes. The proposed model has been demonstrated and explained using a simple example. We analyzed the existing active digital image forensic model and the proposed model based on different measures like robustness, authenticity, integrity, security, computational complexity, and computational cost. We found that in active digital image forensics all the above measures were optimum when visual cryptography scheme was used. Our future study will focus to develop

Table 1 Analysis of existing and proposed model in digital image forensics

Measures	Authentication and tampering detection methods		
	Digital watermarking	Digital signature	Visual cryptography
Robustness	Low	Low	High
Authenticity and integrity	Low	Low	High
Security	Low	Low	High
Computational complexity	High	High	Nil (When using XOR, very low)
Computational cost	High	High	Nil (When using XOR, very low)

more sophisticated models in active digital image forensics using various visual cryptography schemes for grayscale and color images. Variations of this method are also possible for various needs.

References

1. Redi, J.A., Taktak, W., Dugelay, J.-L.: Digital image forensics: a booklet for beginners. Multimed Tools Appl. **51**(1), 133–162 (2011). https://doi.org/10.1007/s11042-010-0620-1(Springerlink.com)
2. Carvalho, T., Faria, F.A., Pedrini, H., Torres, R.D.S., Rocha, A.: Illuminant-based transformed spaces for image forensics. IEEE Trans. Inf. Forensics Security **11**(4), 720–733 (2016). (IEEE) https://doi.org/10.1109/tifs.2015.2506548
3. Warbhe, A.D., Dharaskar, R.V., Thakare, V.M.: Computationally Efficient Digital Image Forensic Method for Image Authentication. Procedia Computer Sci. **78**, 464–470 (2016). (Elsevier), https://doi.org/10.1016/j.procs.2016.02.089
4. Ardizzone, E., Bruno, A., Mazzola, G.: Copy-move forgery detection by matching triangles of keypoints. IEEE Trans. Inf. Forensics Security **10**, 2084–2094 (2015). (IEEE) https://doi.org/10.1109/tifs.2015.2445742
5. Bahrami, K., Kot, A.C., Li, L., Li, H.: Blurred image splicing localization by exposing blur type inconsistency. IEEE Trans. Inf. Forensics Security **10(**5), 999–1009 (2015). (IEEE) https://doi.org/10.1109/tifs.2015.2394231
6. Naor, M., Shamir, A.: Visual Cryptography, Advances in Cryptology-Eurocrypt'94. LNCS 950, pp. 1–12 (1995)
7. Monoth, T., Anto, B.P.: Analysis and design of tamperproof and contrast-enhanced secret sharing based on visual cryptography schemes. Ph.D thesis, Kannur University, Kerala, India, (2012). http://shodhganga.inflibnet.ac.in
8. Hwang, K.-F.: Chang, C.-C.: Recent development of visual cryptography. In: Intelligent Watermarking Techniques, Chap. 16, pp. 459–479, (2004) World Scientific
9. Singh, M.N. Joshi, S.: Digital image forensics: progress and challenges. In: Proceedings of 31st National convention of Electronics and Telecommunication Engineers, Researchgate (October 2015)
10. Singh, S., Bhardwaj, R.: Image Forgery detection using QR method based on one dimensional cellular automata. In: Ninth International Conference on Contemporary Computing (IC3) 2016, IEEE Explore (2017). https://doi.org/10.1109/ic3.2016.7880197

11. Komal, S., Mehta, A.K.: A review on digital watermarking techniques. Appl. Attacks. IJECS **4**, 11237–11245, 015
12. Monoth, T., Babu, A.P.: Tamperproof transmission of fingerprints using visual cryptography schemes. Procedia Comput. Sci. **2**, 143–148 (2010). (ELSEVIER)
13. Tyagi, S., Singh, H.V., Agarwal, R., Gangwar, S.K.: Digital watermarking techniques for security applications. In: International Conference on Emerging Trends in Electrical, Electronics and Sustainable Energy Systems (ICETEESES–16), IEEE Explore (2016). https://doi.org/10.1109/iceteeses.2016.7581413
14. Guojuan, Z., Dianji, L.: An Overview of digital watermarking in image forensics. In: Fourth International Joint Conference on Computational Sciences and Optimization (CSO), IEEE Xplore (2011). https://doi.org/10.1109/cso.2011.85
15. Yadav, U., Sharma J.P., Sharma, D., Sharma, P.K.: Different watermarking techniques & its applications: a review. Int. J. Sci. Eng. Res. **5**(4) (2014)
16. Singh, P,. Chadha, R.S.: A Survey of digital watermarking techniques, applications and attacks. Int. J.Eng. Innovative Technol. (IJEIT), **2**(9) (March 2013)
17. Durvey, M., Satyarthi, D.: A review paper on digital watermarking. Int. J. Emerging Trends & Technol. Comput. Sci. (IJETTCS) **3**(4) (2014)
18. Kusuma Kumari, B.M.: A Survey of digital watermarking techniques and its applications. Int. J. Sci. Res. (IJSR), **2**(12) (2013)
19. von Wangenheim, A., Custódio, R.F., Martina, J.E., de Back Giuliano, I., Andrade, R.: Digital signature of medical reports: an issue still not resolved. Revista da Associação Médica Brasileira (English Edition), **59**(3), 209–212 (January 2013). (ELSEVIER)
20. Liu, C.: Security analysis of liu-zhang-deng digital signature scheme. In: 5th International Conference on Ambient Systems, Networks and Technologies (ANT-2014), Procedia Computer Science 32, pp. 485–488 (2014). (ELSEVIER)
21. Heckeroth, J., Boywitt, C.D.: Examining authenticity: an initial exploration of the suitability of handwritten electronic signatures. Forensic Sci. Int. **275**, 144–154 (2017). (ELSEVIER)
22. Fernández-Alemán, J.L., Señor, I.C., Lozoya, P.Á.O., Toval, A.: Security and privacy in electronic health records: a systematic literature review. J. Biomed. Inf. **46**, 541–562 (2013). (ELSEVIER)
23. Hodeisha, M.E., Bukauskasb, L., Humbe, V.T.: An Optimal (k, n) visual secret sharing scheme for information security. In: 6th International Conference On Advances In Computing & Communications, ICACC 2016, 6–8 September 2016, Cochin, India Procedia Computer Science, vol. 93, pp. 760–767,2016. (ELSEVIER)
24. Dahata, A.V., Chavan, P.V.: Secret sharing based visual cryptography scheme using CMY color space. Procedia Comput. Sci. **78**:563–570 (2016). (ELSEVIER)

Studying the Contribution of Machine Learning and Artificial Intelligence in the Interface Design of E-commerce Site

Megharani Patil and Madhuri Rao

Abstract The impact of machine learning and artificial intelligence areas in e-commerce is growing. Algorithms from these areas help to grow sales and optimize various aspects of e-commerce operation, right from product selection to successful ordering of products. This work is focused on recommender system, navigation optimization, and product review summarization using machine learning and artificial intelligence techniques. Demographic content-based collaborative recommendation system framework is designed using hybrid similarity measure. Navigation optimization is done using the optimized prefix span algorithm. Gibbs sampling based latent Dirichlet allocation classifier framework is used to classify product reviews into positive, negative, and neutral, and represents it in bar chart form. These contributions will reduce human efforts while shopping using e-commerce site and helpful for high-quality user experience with more relative efficiency and satisfaction level.

1 Introduction

Artificial intelligence and machine learning help the e-commerce to engage with their customers on a new level and create interface design easy to learn, efficient to use, and pleasant for better user experiences. As e-commerce industry scales up, these areas made it easier to stay in the game. Also, it makes e-commerce marketers' lives easier. A recommender system is the information retrieval method using which users can make decisions and overcome information overload. It helps users to make decisions in an e-commerce scenario either for product selection or to rate newly purchased product. A user would obviously prefer a website that recommends him something that is useful to him over a website that simply requires users to navigate into the site to find the products the user need. Predicting a user's preferences by

M. Patil (✉) · M. Rao
Information Technology Department, Thadomal Shahani Engineering College, Mumbai, India
e-mail: megharanitpatil@gmail.com

M. Rao
e-mail: my_rao@yahoo.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_20

measuring similarity with another group of users is the core of a recommender system. Navigation optimization is introduced for faster shopping by optimizing navigation path, i.e., by reducing clicks. The user behavior can be known from the actual usage patterns. The usage patterns extracted from weblogs are recorded from operational websites. This can be done by processing the log data which consist of the task-oriented transactions and applying a sequential mining algorithm to discover frequent patterns from transaction sequence. These frequent patterns will help to identify the usability issues and also suggest the corrective measures to improve it. Review summarization makes product learning easier, and reading each and every review related to the product is not feasible. Consumers are referring other consumers' online reviews while making their shopping decisions. There are the bulk of product reviews which makes browsing a large number of reviews, and learning information of interest is time-consuming and difficult to summarize naturally.

2 Related Work

Recommendations help users to predict the ratings in e-commerce easily with fewer efforts and time. Also, it becomes easy to select a movie in future by viewing a list of recommendations which enhances intuitiveness of user interfaces. Collaborative filtering recommendation algorithm recommends items to similar users based on their preferences. Predicting preferences through user's correlated subgroups produces better results. Subgroup-based collaborative filtering algorithm improved top N recommendations but it has less recommendation accuracy [1]. Typicality-based collaborative filtering finds user's neighbor based on user typicality degrees in user groups which increases accuracy [2]. Weight-based item recommendation approach provides recommendation accuracy without increasing computational complexity [3]. To group products into distinct clusters, a self-constructing clustering algorithm is applied. Recommendation algorithm is applied to resulting clusters by recommending products to each user [4]. Content-based filtering recommends items with same features of items which user's purchased previously [5]. It has the limitation of overspecialization, i.e., user is limited to being recommended similar items already purchased [6]. Navigation optimization provides easy and quick shopping process which enhances intuitiveness of user interfaces. Literature focuses on applying web usage mining on server log records using Apriori and FP tree algorithm [7]. Apriori algorithm requires more time for candidate creation. Generalized Sequential Pattern (GSP) algorithm is used which mines the frequent sequences of terms to find out interests of users. The system performance and user experience are improved by analyzing the log file [8]. In GSP algorithm, more time is required for candidate sequence generation by scanning database iteratively. In the conditional sequence-mining algorithm, the access patterns are used from compact tree structures which are used for generating recommendations [9]. Reading every review for each product is time-consuming and cognitively challengeable. Review summarization decreases user's cognitive load and makes interfaces more intuitive using a popular method like Latent Dirichlet Allocation (LDA). It is a topic modeling method that allows classifying sentences of reviews under a different topic. It has two steps: probability estimation and inference [10]. In literature, a naïve Bayes based latent Dirichlet allocation is used to model the feature space. It uses latent topics as the features to reduce feature dimension to increase accuracy [11]. Topics are a probability distribution over words. Topic modeling is used to identify review aspects by preprocessing movie review dataset, calculating TF-IDF, and giving a result to LDA model [12].

3 Demographic Content-Based Collaborative Recommendation System Framework

This work proposes a framework for a recommendation which takes care of user's demographic information, user's preferences, and user behavior to predict ratings for nonrated items. Also, recommend a list of top N relevant products to users. Rating prediction is done using a hybrid similarity measure which is more accurate than Cosine and Pearson similarity measure. It is compared with Cosine and Pearson on basis of parameters Mean Absolute Error Rate (MAE), precision, recall, and F1 score. This work considered dataset from MovieLens (www.movielens.umn.edu) with 943 users, 100,000 ratings for 1682 movie titles with 18 genres. Each user has rated at least 20 movies. The dataset is randomly divided into training set and test set with the ratio of 80 and 20%. Make use of the training set to train our recommendation model, and then apply the trained model on the test set to make a prediction, and finally compare the difference between predicting ratings and actual ratings using MAE, precision, recall, and F1-score represented in Eqs. (1)–(4).

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \bar{y}_j\right| \tag{1}$$

$$\text{precision} = \text{(float) truePositives / (truePositives + falsePositives)} \tag{2}$$

$$\text{recall} = \text{(float) truePositives / (truePositives + falseNegatives)} \tag{3}$$

$$\text{f1 Score} = \text{(float) } (2 * \text{recall} * \text{precision}) / (\text{recall} + \text{precision}) \tag{4}$$

Table 1 represents hybrid method that has less MAE, more precision compared to Cosine similarity measure, and Pearson similarity measure.

Demographic content-based collaborative recommendation system framework has three main steps:

1. User's clustering based on user's demographic information: Apply K-mean clustering algorithm to divide users into K clusters based on their demographic information age, gender, occupation, and zip code.
2. Rating prediction using hybrid similarity measure: Using a hybrid similarity measure, i.e., hybrid of Pearson correlation similarity and Cosine similarity considering hybrid ratio (= 0.5) [13], predicted ratings are suggested to user for

Table 1 Comparison of similarity measures on basis of MAE, precision, recall, and F-score

Parameters	Cosine	Pearson	Hybrid
MAE	3.52619976	3.52619969	3.52619524
Precision	0.29166666	0.3125	0.32584271
Recall	0.00070	0.00075	0.00145
F-score	0.00139665	0.00149641	0.00288715

product which will help user to rate newly purchased product more accurately and to get appropriate product recommendations. Equation (5) represents formulas for cosine similarity measure between two users, x and y.

$$\cos(\theta) = \frac{\sum x_i * y_i}{\sqrt{\sum x_i^2} * \sqrt{\sum y_i^2}} \tag{5}$$

Equation (6) represents Pearson similarity measure computes similarity by subtracting the user's average rating value based on cosine similarity calculation.

$$p_{xy} = \frac{\sum((x_i - \bar{x}) * (y_i - \bar{y}))}{\sqrt{\sum(x_i - \bar{x})^2} * \sqrt{\sum(y_i - \bar{y})^2}} \tag{6}$$

The hybrid similarity measure is given by Eq. (7).

$$H = \partial * p_{xy} + (1 - \partial) * \cos(\theta) \tag{7}$$

Predicted rating for the xth user for product i is given by Eq. (8), where y is the neighbor of x.

$$\bar{r}_{x,i} = \bar{r}_x + \frac{\sum_{y \in N(x)} \mathrm{sim}(x, y) * (r_{y,i} - \bar{r}_y)}{\sum_{y \in N(x)} |\mathrm{sim}(x, y)|} \tag{8}$$

3. Product recommendations using content-based collaborative filtering.
   - 3.1 Calculate product feature matrix P from user feature and behavior matrix U and R.
     - 3.1.1 Formation of set of polynomial equation $U_i^T P_j = R_{ij}$ by referring user feature matrix and behavior matrix.
     - 3.1.2 Solving polynomial equation, product vector for each item is calculated and item feature matrix is formed.
   - 3.2 Product–user mapping vector is calculated by equation MAX $(U_{(j)}^T. I_{(i)})$.
   - 3.3 Top N products are recommended by referring item–user mapping vector. To check real user experience, 100 users have newly registered and each one has rated at least 20 movies with the same system. A web page is created with the suggested rating and a recommended list of movies for a

Table 2 Average time taken to rate movie

Approach	Time required (s)
With recommendation	5
Without recommendation	15

Table 3 Satisfaction level of users who have used recommender system

Questionnaire	Satisfaction level
Q1. I found rating with this interface is easy	4.7
Q2. I found rating with this interface is fast	4.2
Q3. The list contains movies I am thinking for	4.4
Q4. This list is personalized for me	4.2
Q5. The level dissimilarity among the recommended movies is right for me	3.7
Q6. The level of popularity of recommended movies is right for me	4.2
Q7. The level of serendipity about the movies is right for me	4.33
Total satisfaction level	4.25

user. To check user experience, task-based usability testing is carried out by above 100 users. They rated movies watched by them in both scenarios. In website version without recommender system it takes more time to rate movie than the website version with the added recommender system as shown in Table 2.

User satisfaction level for a website with recommender system [14] is measured by setting seven questionnaires. Table 3 represents satisfaction level of users who have used recommender system that is on an average 4.25 out of 5 which means they are satisfied with the system.

4 Navigation Optimization with Optimized Prefix Span Algorithm

Prefix span finds the frequent pattern of length 1 and generates its projected databases which are mined separately. The algorithm finds prefix patterns, which is further appended with suffix patterns to find frequent patterns, without candidate sequence generation. Optimized prefix span algorithm is mentioned below:

Optimized Prefix Span Algorithm
Input: A database(S), Threshold minimum support (min_support).
Output: A set of sequential patterns.

Table 4 Time/space complexity for prefix span and optimized prefix span

Algorithm	Time required (ms)	Space required (bytes)
Prefix span algorithm	590	533
Optimized prefix span algorithm	357	309

Data Definition: α: sequential pattern, *L*: length of α and *S*|α: α—projected database if α ≠ <>, Otherwise, it is the sequence database S.

Function Mody_Prefix(<>, 0, *S*)

Step1: Scan $S|\alpha$ once, find each frequent item x, such that x can be assembled to the last element of α to form a sequential pattern.
Step2: For each frequent item x, append it to α to form a sequential pattern α'. If x is not matching with α, go to step 3 otherwise step 4.
Step3: For each α', if support $(\alpha') >= $ min_support, then construct α'-projected database and obtain α′-suffix sequence.
Step4: For each α', if support $(\alpha') >=$ min_support, then $S|\alpha' =$ Sequence database S and obtain α′-suffix sequence.

Call Function Mody_Prefix (α', $L+1$, $S|\alpha'$).
The website is created for virtual electronics products online shop. Totally, 110 web usage transaction patterns of 72 users were collected. These web usage patterns are mined using prefix span and optimized prefix span algorithm. It is found that optimized prefix span required less time and space than prefix span which is listed in Table 4.

The frequent usage patterns extracted were analyzed to identify some usability issues. Accordingly, changes are done on the website which is listed in Table 5. Optimized prefix span is also applied to find frequent patterns from the modified website. 106 web usage transactions of 72 users are extracted for the modified website, and the optimized prefix span algorithm is applied. It is observed that frequent patterns of less path length were obtained in a modified website than original website [15]. As shown in Table 5, average efforts are reduced by 22.62% using modified website which means it is more intuitive.

5 Review Summarization Using Gibbs Sampling Based Latent Dirichlet Allocation Classifier Framework

This work introduces latent Dirichlet allocation classifier framework that derives a procedure which first does probability estimation followed by that topic assignment as given below:

Table 5 Frequent pattern for original and modified web site

Original website		Modified website		% Efforts reduced (%)	Changes done in modified website
Frequent patterns	Length	Frequent patterns	Length		
Home, appliances, tv, buy	4	Home, LG, buy	3	25	Popular products are displayed on homepage with "buy" button
Home, laptop, macbookair, home, laptop, mackbookpro, buy	7	Home, laptop, hp, Lenovo, buy	5	28.57	Back button between product page and catalog page
Home, camera, Nikon, home, mobile, OnePlus, buy	7	Home, camera, Canon, mobile, Samsung, buy	6	14.29	Display of submenus for categories is done on category page

1. Preprocessing list of reviews: Remove all stop words from reviews and perform stemming and lemmatization of each word from reviews. Also, create a dictionary with positive and negative words.
2. Apply LDA for a document D consisting of M reviews each of length $N_{i.}$

   2.1 Choose $\Theta_{i\,\sim}$ Dir(α) where $i \in \{1 \ldots M\}$ and Dir(α) is a Dirichlet distribution with a parameter $\alpha < 1$.

   2.2 Choose $\varphi_{k,}$ distribution of words in topic k ~Dir(β) where $k \in \{1 \ldots . k\}$ and Dir(β) is a Dirichlet distribution with a parameter α which is much less than 1.

   2.3 For each of the word positions i, j, where $j \in \{1 \ldots N_i\}$ and $i \in \{1 \ldots M\}$. Choose a topic $z_{\,i,j}$ ~Multinomial($\Theta_{i)}$ and choose a word $w_{i,j}$ ~Multinomial(φ $z_{i,j).}$

3. Probability estimation and inference of topic assignment for each of word w, a multimodal probability conditioned on the topic using Gibbs sampling. Gibbs sampling updates the hard assignment z_i of a word token w_i to one of the topics $k \in \{1 \ldots K\}$. This update is performed sequentially for all word tokens from each review; for a fixed number of iterations, n_{kv} represents the number of times that word type v is assigned to topic k across the reviews, n_{dk} represents the number of word tokens in document d assigned to topic k, and $\neg i$ represents the current topic assignment of w_i. The probability of setting topic assignment z_i to topic k, conditional on w_i, d, the hyperparameters α and β is given in Eq. 9 [16].

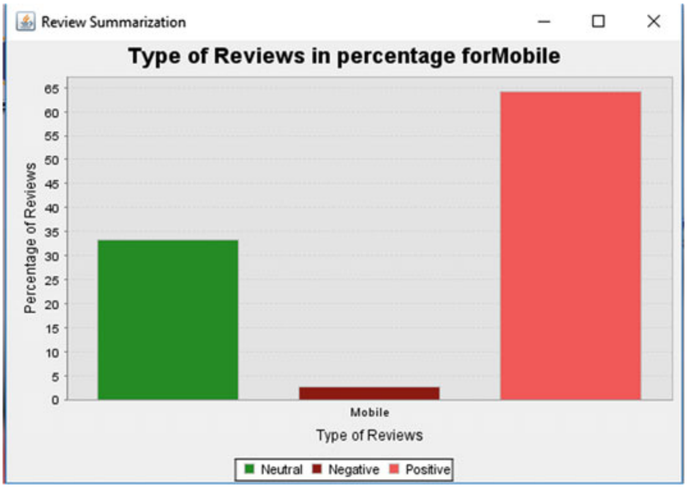


Fig. 1 Review summarization into positive, negative, and neutral

Table 6 Comparison of accuracy and time span

LDA	Accuracy (%)	Time span (*10^3 s)
Primitive variational interference based	62.85	5.4
Gibbs sampling based	77.4	3.9

$$\rho = (z_i = k | w_i = v, d, \alpha, \beta, .) \propto \frac{n_{kv\neg i} + \beta_v}{\sum_{v'=1}^{V} (n_{kv'\neg i} + \beta_{v'})} . \frac{n_{dk\neg i} + \alpha_k}{N_d + \sum_{k'=1}^{K} \alpha_{k'}} \tag{9}$$

4. A bar chart is formed by a count of positive, negative, and neutral review count from the input document.

Latent Dirichlet allocation classifier framework is applied to Amazon product review dataset available for Alcatel Smartphone with total 907 reviews. Figure 1 shows review summarization in bar chart form with the percentage of positive: 64.10%, negative: 2.5%, and neutral: 33.33% in graphical form.

Gibbs sampling based latent Dirichlet allocation classifier framework has more accuracy and requires less time in comparison with primitive variational inference based latent Dirichlet allocation as shown in Table 6.

To check user experience, task-based usability testing is carried out by 100 users. Table 7 shows that the time required to learn descriptive reviews is more than to learn summarized reviews in graphical form.

To measure user's satisfaction level [17] and to learn reviews with review summarization using bar chart survey are conducted by setting five questionnaires. Table 8

Table 7 Time required to learn descriptive and summarized reviews

Approach	Time required (s)
Descriptive reviews	50
Summarized reviews	4

Table 8 Satisfaction level of users who have referred summarized reviews

Questionnaire	Satisfaction level
Q.1 I found the task was easy	4.2
Q.2 I required fewer efforts to complete the task	4.6
Q.3 I felt successful in accomplishing the task	3.7
Q.4 I was satisfied with the amount of time it took me to complete the task	4.3
Q.5 I was not discouraged, irritated while doing the task	4.4
Total satisfaction level	4.24

shows that average user satisfaction level is 4.24 out of 5, indicating users are satisfied with summarized reviews.

6 Conclusion and Future Work

Here, we have represented contribution for features such as machine learning and artificial intelligence by adding demographic content-based collaborative recommendation system framework, navigation optimization through optimized prefix span algorithm, and review summarization using Gibbs sampling based latent Dirichlet allocation classifier framework modules which have reduced human efforts and has increased user satisfaction level. In this way, machine learning and artificial intelligence have contributed to designing intuitive interfaces for e-commerce shopping site.

References

1. Bu, J., Shen, X., Xu, B., Chen C., He, X., Cai, D.: Improving collaborative recommendation via user-item subgroups. IEEE Trans. Knowledge Data Eng. **6**(1) (2007)
2. Cai, Y., Leung, H., Li, Q., Min, H., Tang, J., Li, J.: Typicality-based collaborative filtering recommendation. IEEE Trans. Knowledge Data Eng. (2013)
3. Zhao1, Y., Liu, Y., Zeng, Q.: A weight-Based Item Recommendation Approach for Electronic Commerce Systems. Springer (2015)

4. Liao, C., Lee, S.: A Clustering based approach to improving the efficiency of collaborative filtering recommendation. Electron. Commerce Res. Appl. May 7 (2016)
5. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans. Knowl. Data Eng. **17**(6), 734–749 (2005). https://doi.org/10.1109/TKDE.2005.99
6. Li, S., Karahanna, E.: Journal of the association for information systems online recommendation systems in a B2C e-commerce context : a review and future directions online recommendation systems in a B2C E-commerce context : a review and future directions, **16**(2), 72–107 (2015)
7. Gupta, A., Arora, R., Sikarwar, R., Saxena, N.: Web usage mining using improved frequent pattern tree algorithms. In: 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT). Ghaziabad (2014)
8. Tang, Y., Tong, Q., Du, Z.: Mining frequent sequential patterns and association rules on campus map system. In: The 2014 2nd International Conference on Systems and Informatics (ICSAI 2014). Shanghai (2014)
9. Xiao-Gang, W., Yue, L.: Web mining based on user access patterns for web personalization. In: ISECS International Colloquium on Computing, Communication, Control, and Management. Sanya (2009)
10. Bhagat, T., Patil, M.: Predicting user preference for movies using movie lens dataset. Int. J. Recent Trends Eng. Res. ISSN (Online) **3**(2), 2455–1457, pp. 156–163 (2016). https://doi.org/10.23883/ijrter.2017.3018.k3lx
11. Mahyavanshi, N., Patil, M., Kulkarni, V.: Enhancing web usability using user behavior and cognitive study. Int. J. Comput. Appl. (0975–8887), **164**(2), 27–31 https://doi.org/10.5120/ijca2017913594 (2017)
12. Liu, P.Y., Gong, W., Jia, X.: An improved prefixspan algorithm research for sequential pattern mining. In: IEEE International Symposium on IT in Medicine and Education, Cuangzhou (2011)
13. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. J. Machine Learning Res. 3, ISSN: 993–1022 (2003)
14. Chang, Y., Chien, J.: Latent dirichlet learning for document summarization. In: IEEE International Conference On Acoustics, Speech And Signal Processing 978-1-4244-2354-5/09/ (2009)
15. Mohana, R., Umamaheswari, K., Karthiga, R.: Sentiment classification based on latent dirichlet allocation. Int. J. Comput. Appl. (0975–8887) (2015)
16. Nguyen, T.: Enhancing user experience with recommender systems beyond prediction accuracies. In: A thesis submitted to the faculty of the graduate school of the university of Minnesota (2016)
17. Mifsud, J.: Usability metrics—a guide to quantify the usability of any system. https://usabilitygeek.com/usability-metrics-a-guide-to-quantify-system-usability/ [Accessed on Nov 2017] (2015)

Discrete Wavelet Transform Based Invisible Watermarking Scheme for Digital Images

Swapnil Satapathy, Khushi Jalan and Rajkumar Soundrapandiyan

Abstract Authenticity is one of the major concerns for data privacy in the world at present. There must be ample techniques available to the user in order to protect their data from piracy or copyright. One such technique to protect a user's data from piracy is digital watermarking. Digital watermarking is the process in which a marker is covertly embedded in signals that are noise tolerant such as audio, video, or image data. In this paper, Discrete Wavelet Transform (DWT)-based invisible watermarking scheme is proposed for digital images. In the proposed method, the embedding factors are calculated using kurtosis and skewness. The proposed method can be easily utilized for ownership of copyright of that particular image. The proposed method is powerful and makes sure that the effect of any signal and image processing attacks on the watermark is minimal. The experimental analysis and results are provided along with statistical analysis to show the efficiency and accuracy of this proposed technique. The contributions of the proposed method are as follows: (1) It can be used for images of any size and scale. (2) It is very simple and the time taken for processing the image is very less. (3) It has been scrutinized under various attacks and has been verified.

1 Introduction

In recent years, there have been a lot of discoveries and innovations which today has lead to the increase in the amount of information which in turn has given rise to various problems that are becoming very difficult to be dealt with. This situation that is persisting today is termed as information explosion. Information explosion can be defined as the vast amount of increase in the amount of information that is published and the abundant data that is available which leads to huge problems in managing the information available. This can lead to information overload. Large amounts of information are available easily on the Internet to the users. The Internet

S. Satapathy · K. Jalan · R. Soundrapandiyan (✉)
School of Computer Science and Engineering, VIT, Vellore 632014, India
e-mail: rajkumarsrajkumar@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_21

can be considered as a boon for the users, but it also comes out to be a bane for the actual owners of the information and data that is available on the Internet is used by the end users. It affects the actual owners as there is a huge amount of unauthorized and copyright violated information that is easily available on the Internet. So in order to tackle this serious problem of providing legal ownership rights and authenticity to the actual owners of the information, the data must be encrypted and secured using suitable techniques like cryptography, steganography, watermarking, etc. Digital watermarking is one such technique that is used to protect the ownership rights. This method basically works on the principle of information hiding that is broadly used to prevail over illegal copyright of digital information.

Digital watermarking is a technique that helps to identify copyrighted information. It hides the digital information in a carrier signal. In the process of digital image watermarking, a digital text or logo is embedded in an invisible domain so that it cannot be tampered. This leads to the protection of the data and helps maintain its authenticity and integrity. Illegal distribution of data can be largely prevented using this technique of digital watermarking. To carry out the process of digital image watermarking, image processing techniques are extensively used. The watermarking system is incorporated in two steps: watermark embedding and watermark extraction. There are different methods in which these steps can be achieved. The process of embedding the watermark can be done in either a spatial domain or the frequency domain. The embedding of watermark in the spatial domain is the technique in which the pixels of the original image are tampered with the pixels of the watermarked image, whereas the embedding of watermark in frequency domain is done using the concept of the spectrum. The watermark can be embedded into the coefficients of Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT), or Discrete Fourier Transform (DFT). This technique of embedding the watermark in frequency domain results in the increase of robustness when compared to the spatial domain.

The main features to be taken into considering while designing the watermark are robustness, indiscernibility, and dependability. The watermark must be robust and should not degrade while transformations. It must be indiscernible, that is, it must not be noticeable when compared to the original image. The image must be perfectly similar and unnoticeable with the original image. The watermark must be dependable and must not degrade. These are the desirable features that must be taken into account in the making of the watermark. The features play the important role of ensuring the security and prevent from attacks.

The digital image watermarking when done in DWT definitely has given better results and holds the upper hand when compared to the other frequency domains. There are several research works done in the recent past to improve the characteristics of digital watermarking. Su and Chen [1] analyzed the upper Hessenberg matrix and proposed a blind color image watermarking. Najih et al. [2] proposed a watermarking scheme based on angle quantization in discrete contourlet transform. Lai and Tsai [3] proposed a hybrid image watermarking method that used DWT and singular value decomposition method to improve the robustness of the watermark. Purwar and Jain [4] also proposed a digital image watermarking method that involved DWT and SVD, but in addition to that, a multifactor scale was used. A three-level DWT was performed after which singular value decomposition was applied. Purohit et al. [5] applied a different technique that used single-level Stationary Wavelet Transform (SWT) which also involved Singular Value Decomposition (SVD) which improved the characteristics of Tiwari et al. [6] who presented novel digital watermarking technique that involved the use of DWT and Fast Fourier Transform (FFT) along with SVD. This technique also was used to increase the robustness of the watermark. Mudassar et al. [7] presented a technique for digital image watermarking that involved a hybrid DWT-SVD technique.

In this paper, an invisible watermarking technique is proposed that uses DWT. The factors that are used for embedding the watermark are skewness and kurtosis. The results show a promising output that is needed to increase the robustness of the watermark.

The rest of the paper is organized as follows: In Sect. 2, the proposed method is briefly discussed. The experimental results are presented in Sect. 3. Section 4 concludes the work.

2 Proposed Method

The proposed method works in two stages, namely, embedding and extraction. The block diagram of the proposed method is shown in Fig. 1.

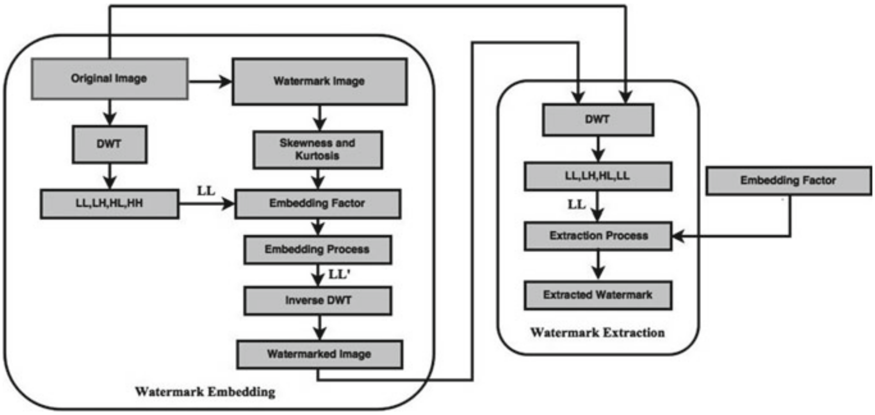


Fig. 1 The block diagram of the proposed method

2.1 Watermark Embedding

The process of embedding a watermark in a cover image is done in three main steps. The three steps are

1. Discrete wavelet transform,
2. Calculation of embedding and scaling factors, and
3. Embedding.

2.1.1 Discrete Wavelet Transform

DWT is a multi-resolution analysis, which decomposes an image into wavelet coefficients and scaling function. It is performed by passing an image through low-pass and high-pass filtering. In doing so, the given input image decomposes into a low-frequency band which is also the approximation subband (LL), two middle-frequency subbands (LH, HL) containing the horizontal and vertical details of the image, and a high-frequency band (HH) containing the diagonal details. Among the various types of wavelets, in this paper, Haar wavelet is used which is the simplest type of wavelet. Here, the approximate bands are analogous to the low-frequency components while the detail bands are analogous to the high-frequency components of the wavelets.

The wavelets obtained from DWT act as base functions for many signal and image representations. This method of transformation is a very significant way of transformation which converts an image from the spatial domain to it its corresponding frequency domain, which thereby allows us to make simultaneous interpretations for both the domains. This significant characteristic of DWT is widely used as it increases the indistinguishability of the watermark image.

2.1.2 Calculation of Embedding and Scaling Factors

The two main variables that are used to calculate the embedding and scaling factors for embedding the watermark in a cover image are kurtosis and skewness.

Kurtosis is the descriptor of the tailedness (shape) of the probability distribution of a random variable. Hence, it is used as an embedding and a scaling factor. It is defined in Eq. (1)

$$k = \frac{\mu_4}{\sigma_4} \tag{1}$$

where σ is the standard deviation and μ_4 is the fourth central moment.

Skewness is the measure of the asymmetry or the unevenness of the shape of the probability distribution of a random variable. It can have both positive and negative values. It is the third standardized moment and is defined in Eq. (2)

$$\gamma_1 = \frac{\mu_3}{\sigma_3} \tag{2}$$

where σ is the standard deviation and μ_3 is the third central moment. Further, to get the exact shape of the distribution of the image, sigmoid function is used. It is defined in Eq. (3)

$$l = \frac{1}{1 + \mathrm{e}^{-t}} \tag{3}$$

where t is the skewness.

Kurtosis being the shape descriptor of the probability distribution of a random variable provides shape to its graph which may not be perfectly bell-shaped for all the values. This is the reason why skewness has been used to provide the proper shape of the probability distribution graph of that particular random variable.

The scaling and embedding factors are represented by *a* and *b,* respectively, which are defined in Eqs. (4) and (5).

$$a = k - l \tag{4}$$

$$b = 1 - k - l \tag{5}$$

where k and l are the kurtosis and sigmoidal of skewness values, respectively.

2.1.3 Embedding

The insertion of the watermark in the cover image is carried out in the following steps:

Step 1 Read the cover image (A) and watermark image (B).

Step 2 Calculation of scaling and embedding factors a and b

Step 3 The cover image (A) decomposed into LL, LH, HL, and HH subbands by applying level 2 DWT using Haar wavelet.

Step 4 Extraction of LL band from the wavelets obtained.

Step 5 Insertion of watermark in the LL band using Eq. (6)

$$\mathrm{LOW}' = a \times \mathrm{LOW} + b \times \mathrm{WM} \tag{6}$$

where LOW is the LL band of the cover image, LOW′ is the modified LL band, and WM is the watermark image.

Step 6 Combining the LOW′ band with the LH, HL, and HH bands obtained in step 3.

Step 7 Inverse DWT is applied to get watermarked image.

2.2 Watermark Extraction

The process of extraction of watermark from a watermarked image is exactly the reverse of the process of embedding the watermark. The steps carried out for the extraction are as follows:

Step 1 Read the cover image (A) and watermarked image (C).
Step 2 Get the scaling and embedding factors calculated from the embedding process.
Step 3 Application of level 2 DWT on the cover image and watermarked image using Haar wavelet.
Step 4 Extraction of LL band from the wavelets obtained on both the images.
Step 5 Extraction of watermark from the LL band using Eq. (7):

$$\mathrm{WM} = \frac{\mathrm{LOWM} - a \times \mathrm{LOW}}{b} \tag{7}$$

where WM is the extracted watermark, LOWM is the LL band of the watermarked image, and LOW is the LL band of the cover image.

3 Experimental Results and Performance Analysis

To show the performance of the proposed method, several experiments are conducted on a particular set of images to demonstrate the effectiveness of the aforementioned watermarking scheme. The set of 10 different images used in the grayscale format as the cover image for the demonstration are “airplane”, “barbara”, “girl”, “girl2”, “house”, “house2”, “lena”, “mandrill”, “peppers”, and “tree”. The size of the image used is 256 × 256. The image used as the watermark is “cameraman”, and the size of the image used is 64 × 64. The sample cover images and the watermark image are shown in Fig. 2. Figure 3 shows the sample watermarked images. The efficacy of this scheme has been displayed by the Peak Signal-to-Noise Ratio (PSNR) and the Normalized Correlation Coefficient (NCC) values. PSNR values calculated between the watermarked image and the cover image help us to assess the indistinguishability of the watermark and are evaluated by the Mean Square Error (MSE) values. PSNR and MSE [11, 12] are defined in Eqs. (8) and (9), respectively.

$$\mathrm{PSNR} = 10\log_{10}\left(\mathrm{MAX}^2/\mathrm{MSE}\right) \tag{8}$$

$$\mathrm{MSE} = \frac{1}{pq}\sum_{i=0}^{p-1}\sum_{j=0}^{q-1}[A(i, j) - I(i, j)]^2 \tag{9}$$

where MAX is the maximum value of an image. $I(i, j)$ is the value of watermarked image, and A (I, j) is the value of the cover image.

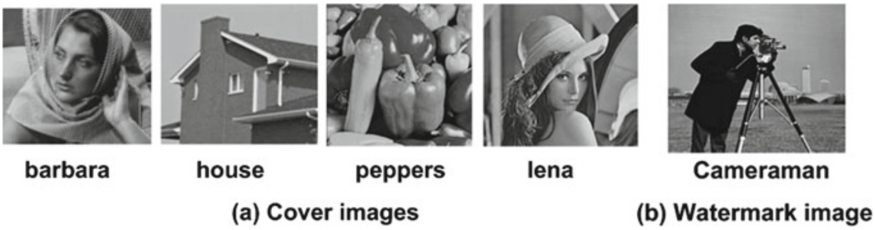


Fig. 2 Sample cover images and watermark images

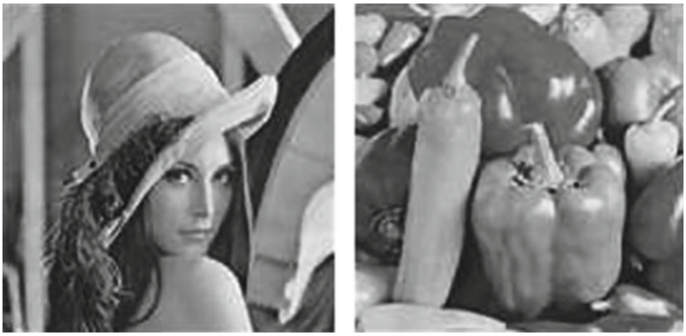


Fig. 3 Sample watermarked images

NCC measures the similarity of the information of two (watermark and extracted watermark) images. It is calculated in Eq. (10)

$$\mathrm{NCC} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} B(i,j) * \mathrm{WM}(i,j)}{\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} B(i,j)^2}\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} \mathrm{WM}(i,j)^2}} \tag{10}$$

where $B(i, j)$ is the watermark image value and $\mathrm{WM}(i, j)$ is the extracted watermark image value. The correlation coefficient ranges vary from −1 to +1. A +1 indicates the positive correlation between the watermark and extracted watermark image. Further, to test the robustness of the proposed method, various types of attacks are applied to the watermarked images like noise, cropping, rotation, and scaling. The retrieved watermark images for the various attacks are shown in Fig. 4. The PSNR values obtained from the proposed method are shown in Table 1.

The NCC values attained from the proposed method are shown in Fig. 5. In addition, PSNR value of the proposed method is compared with the Vaidya et al., Wu et al., and Peng et al. and it is shown in Table 2.

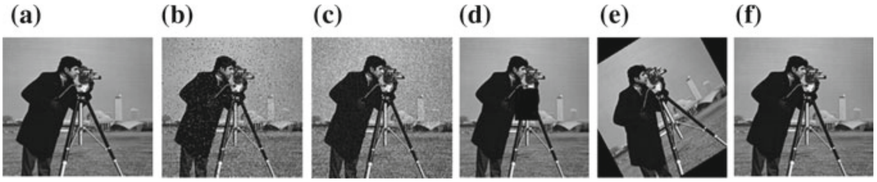


Fig. 4 Retrieved watermark images with **a** no attack, **b** salt and pepper noise, **c** Gaussian noise, **d** cropping, **e** rotation, and **f** scaling

Table 1 Comparison of PSNR values between the cover and watermarked image on various attacks

Cover images	No attack	Salt and pepper noise attack	Gaussian noise attack	Cropping attack	Rotation attack	Scaling attack
Airplane	62.30	60.71	61.37	43.01	63.04	61.29
Barbara	62.07	62.25	62.45	46.16	68.84	62.20
Girl	25.92	26.41	27.02	54.72	27.34	26.49
Girl1	10.35	9.27	8.10	100.36	1.15	9.74
House	90.15	87.09	86.17	144.17	83.33	89.11
House2	43.38	42.22	42.79	64.92	50.97	42.32
Lena	36.58	36.83	36.96	4.31	37.26	37.45
Mandrill	40.31	39.29	38.88	20.70	44.36	39.98
Peppers	68.71	66.74	66.52	100.24	78.97	68.56

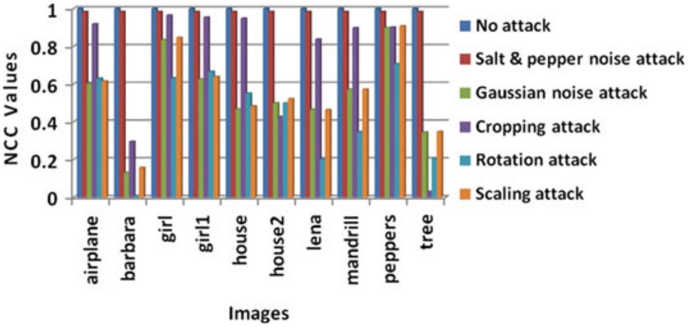


Fig. 5 Comparison of NCC values between the watermark and extracted watermarked image on various attacks

Table 2 Comparison of PSNR values of the proposed method with the existing methods

Images	Peng et al. [9]	Wu et al. [10]	Vaidya et al. [8]	Proposed method
Barbara	29.30	46.89	52.07	62.07

4 Conclusion

In this paper, a novel method for the calculation of embedding and scaling factors was proposed using skewness and kurtosis. The calculated values were used for invisible watermarking in wavelet domain. The proposed method is efficient and robust, which is observed from the subjective and objective experimental results. In future, we would like to extend this watermarking method for copyright protection of the images and also use some machine learning techniques for watermarking colored images.

References

1. Su, Q., Chen, B.: A novel blind color image watermarking using upper Hessenberg matrix. AEU Int. J. Electron. Commun. **78** (2017)
2. Najih, A., Al-Haddad, S.A.R., Ramli, A.R., Hashim, S.J., Nematollahi, M.A.: Digital image watermarking based on angle quantization in discrete contourlet transform. J. KSU Comput. Inf. Sci. **29**(3), 288–294 (2017)
3. Lai, C.C., Tsai, C.C.: Digital image watermarking using discrete wavelet transform and singular value decomposition. IEEE Trans. Instrum. Meas. **59**(11), 3060–3063 (2010)
4. Purwar, R.K., Jain, A.: An evolutionary algorithm based multiscale digital image watermarking technique using discrete wavelet transform and singular value decomposition. Int. J. Tom. Sim. **30**(2), 53–62 (2017)
5. Purohit, N., Chennakrishna, M., Manikantan, K.: Novel digital image watermarking in SWT + SVD domain. In: International Conference on Signal, Networks, Computing, and Systems, pp. 13–23 (2017)
6. Tiwari, N., Hemrajamani, N., Goyal, D.: Improved digital image watermarking algorithm based on hybrid DWT-FFT and SVD techniques. Ind. J. Sci. Technol. **8**(1) (2017)
7. Mudassar, S., Jamal, M., Mahmood, F.S., Shah, N., Tahir, H.B., Malik, H.: Hybrid DWT-SVD digital image watermarking. Int. J. Comput. **26**(1), 105–108 (2017)
8. Vaidya, S.P., Mouli, P.C.: Adaptive digital watermarking for copyright protection of digital images in wavelet domain. Procedia Comput. Sci. **58**, 233–240 (2015)
9. Peng, F., Li, X., Yang, B.: Adaptive reversible data hiding scheme based on integer transform. Sig. Process **92**(1), 54–62 (2012)
10. Wu, H.T., Huang, J.: Reversible image watermarking on prediction errors by efficient histogram modification. Sig. Process **92**(12), 3000–3009 (2012)
11. Bhateja, V., Patel, H., Krishn, A., Sahu, A., Lay-Ekuakille, A.: Multimodal medical image sensor fusion framework using cascade of wavelet and contourlet transform domains. IEEE Sens. J. **15**(12), 6783–6790 (2015)
12. Krishn, A., Bhateja, V., Sahu, A.: Medical image fusion using combination of PCA and wavelet analysis. In: International Conference on Advances in Computing, Communications and Informatics, pp. 986–991 (2014)

Noise Removal in EEG Signals Using SWT–ICA Combinational Approach

Apoorva Mishra, Vikrant Bhateja, Aparna Gupta and Ayushi Mishra

Abstract Electroencephalogram (EEG) represents the electrical activity of the brain recorded by placing several electrodes on the scalp. EEG signals are complex in nature and consist of various artifacts like ocular, muscular, cardiac, etc. The artifacts removal in EEG signals can be majorly modeled by considering it of type Additive White Gaussian Noise (AWGN) in nature. Independent Component Analysis (ICA) is known for its ability to filter out the artifacts from the signal, and hence it is used to rearrange the source signal into two mixtures in a way that the brain signals and the artifacts get separated, although there is a constraint that ICA can only be performed on multichannel signal input. In the present case as the input EEG is single channel, hence, ICA is applied in combination with Stationary Wavelet Transform (SWT) for noise filtering of EEG signals. The quantitative evaluation of proposed approach has been made using Signal-to-Noise Ratio (*SNR*) parameter which depicts satisfactory filtering at varying intensity levels of AWGN.

1 Introduction

An EEG signal is used to denote the electrical neural actions of brain. They have frequency content ranging from 0.01 to 100 Hz that varies from a few μV to 100 μV. Apart from the classic usage in medical fields, EEG has various other applications in neuromarketing, Brain–Computer Interfaces (BCIs), and biometrics. Among the

A. Mishra · V. Bhateja (✉) · A. Gupta · A. Mishra
Department of Electronics and Communication Engineering, Shri Ramswaroop Memorial Group of Professional Colleges (SRMGPC), Lucknow 226028, Uttar Pradesh, India
e-mail: bhateja.vikrant@gmail.com

A. Mishra
e-mail: apoorvamishra3103@gmail.com

A. Gupta
e-mail: aparnag2430@gmail.com

A. Mishra
e-mail: ayushimishra960@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_22

aforesaid applications, the most versatile application is biometrics. By extracting the suitable features, EEG is used in human biometric recognition [1]. Since EEG signals have extremely small amplitudes and thus can easily be polluted by distinct artifacts such as ocular, muscular, cardiac, glossokinetic, and environmental, these artifacts have a perturbing effect on EEG classic bands. It is important to remove these artifacts so that features can be extracted fruitfully. To tackle these artifacts, innumerable methods and techniques have been discovered by different scholars for filtering of noise in the last few decades [2–4]. This comprises conventional filtering techniques which include bandpass filtering [1] as well as the algorithms used for blind source separation like ICA [5]. Also, simply eliminating noisy EEG instants is one of the frequently used ways. But this process involves checking the data manually, spotting noisy sections, and then finally removing those parts. This procedure is tough and results in undesirable information failure when there is a high strength of impurity [6]. A substitute to the aforementioned procedure is to eliminate the artifacts from the data which comprises different methods such as SWT [7], Independent Component Analysis (ICA) [5, 8], etc. ICA is a multichannel approach and it cannot be applied directly to single-channel EEG signal. It is used to extort statistically independent components from a set of measured components. James and Gibson [5] presented a technique which used ICA to expel ocular artifacts from EEG signals. The extracted components were not ordered. Devuyst et al. [9] proposed a modified ICA algorithm used to eliminate ECG noise [10, 11] in EEG or EOG but this method was computationally inefficient. Hyvarinen et al. [12] applied ICA to short-time Fourier transforms of spontaneous EEG but partitioning of impulsive brain actions into source signals was unsuccessful. Since ICA cannot be applied to a single channel, therefore the combinational approach of SWT and ICA is proposed. SWT is used to decompose the signal and ICA is applied. SWT–ICA components are reconstructed back from denoised signal.

The remaining part of this paper is organized as follows. Section 2 describes the proposed EEG signal denoising approach. The methodologies used at each stage are explained in detail in Sects. 2.1, 2.2, and 2.3, respectively. The experiments performed and the achieved results for the reconstructed EEG signals are reviewed in Sect. 3. Finally, Sect. 4 summarizes the concluded work.

2 Proposed EEG Signal Denoising Approach

Noise can be interpreted as a disturbance which affects the signal peaks and results in signal distortion. EEG signals are considered to exhibit chaotic behavior as they are generated by random processes. Addition of Gaussian noise (AWGN) to EEG signal declines the signal quality, and it becomes difficult to interpret its characteristics. Based on the proposed EEG denoising approach, first of all the noisy input EEG signal is decomposed using the SWT. After signal decomposition, the obtained approximate and detailed coefficients are processed using soft thresholding process. Among the various ICA algorithms, fast ICA is preferred for denoising [13] since it increases

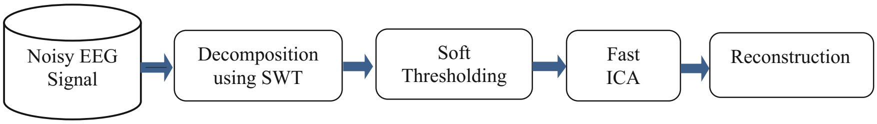


Fig. 1 Block diagram of EEG denoising

the computational efficiency. Finally, reconstruction is done to retrieve the processed denoised signal [2]. The complete methodology has been pictured as per the block diagram in Fig. 1.

Various sub-modules of the above block diagram are explained in the following subsection.

2.1 *Stationary Wavelet Transform (SWT)*

To preserve the translation invariance property, Stationary Wavelet Transform (SWT) is used. In this paper, SWT is chosen for decomposition using mother wavelet "symlet" up to six levels. The decomposition formulae of SWT are shown in Eq. (1).

$$
\begin{aligned}
A_{j,k1,k2} &= \sum_{n1}\sum_{n2} h_o^{\uparrow 2^j}(n_1 - 2k_1)h_o^{\uparrow 2^j}(n_2 - 2k_2)A_{j-1,n_{1,n_2}} \\
D^1_{j,k1,k2} &= \sum_{n1}\sum_{n2} h_o^{\uparrow 2^j}(n_1 - 2k_1)g_o^{\uparrow 2^j}(n_2 - 2k_2)A_{j-1,n_{1,n_2}} \\
D^2_{j,k1,k2} &= \sum_{n1}\sum_{n2} g_o^{\uparrow 2^j}(n_1 - 2k_1)h_o^{\uparrow 2^j}(n_2 - 2k_2)A_{j-1,n_{1,n_2}} \\
D^3_{j,k1,k2} &= \sum_{n1}\sum_{n2} g_o^{\uparrow 2^j}(n_1 - 2k_1)g_o^{\uparrow 2^j}(n_2 - 2k_2)A_{j-1,n_{1,n_2}} \qquad (1)
\end{aligned}
$$

where $A_{j,k1,k2}$, $D^1{}_{j,k1,k2}$, $D^2_{j,k1,k2}$, and $D^3_{j,k1,k2}$ are the low-frequency components, the horizontal high-frequency component, vertical high-frequency component, and diagonal components of the SWT, respectively. $h_o^{\uparrow 2j}$ and $g_o^{\uparrow 2j}$ are used to denote that $2^j - 1$ zeros are inserted between the two points h_o and g_o [14].

2.2 *Soft Thresholding*

Soft thresholding is used for denoising EEG signal [15] by applying it to transform-domain representation of the signal. It shrinks the noisy coefficients above the threshold and makes the algorithms more manageable. After SWT decomposition, soft thresholding is applied to the obtained detailed coefficients. For soft thresholding, the following nonlinear transform is used shown in Eq. (2) [16]:

$$\hat{z}(x) = \text{sign}(x) \cdot (|x| - T_h)_+ \tag{2}$$

Here, x denotes the noisy coefficient, z denotes the noise-free coefficient, and n represents noise. The aim is to estimate w from the noisy observation y. The estimate will be denoted as $\hat{z}$. Because the estimate is dependent on the observed (noisy) value x, estimate can be denoted as $\hat{z}(x)$ as shown in Eq. (3).

$$\hat{z}(x) = \begin{cases} x + T_h,\ x < T_h \\ 0,\ -T_h \ll x \ll T_h \\ x - T_h\ T_h < x \end{cases} \tag{3}$$

$$T_h = \frac{\sqrt{2}\sigma_n^2}{\sigma} \tag{4}$$

where T_h is the threshold.

2.3 Fast ICA

ICA is an approach which extracts statistically independent components from the set of measured signals. ICA algorithm is suitable when the number of sources is greater than the number of channels, i.e., it can be applied to multiple channels only. In the case of single channel, the combination of SWT and ICA has been applied presently. Among various ICA algorithms, fast ICA is used because fast ICA is an efficient and popularly used algorithm for blind source separation [17]. It is computationally efficient and requires less memory over other algorithms as it estimates the independent components one by one. It also has the advantage of multicomponent extraction, and the system performance is not degraded. The modeling of ICA is done by Eq. (5) stated as

$$Z = A_m \cdot s_D \tag{5}$$

where Z denotes the observed matrix, s_D indicates the determined sources, and A_m denotes the separating matrix. ICA is majorly used to identify the separating matrix X so as to attain the independent components under the prerequisites of independent criteria.

$$s_D^* = X \cdot Z \tag{6}$$

$$X = A_m^{-1} \tag{7}$$

Table 1 Various results obtained with varying *SNR* (dB) values using SWT–ICA on signal#1

SNR (dB)	Noisy signal *SNR* (dB)	Reconstructed signal *SNR* (dB)
5	−7.352165	25.296003
10	−7.348412	25.200892
15	−7.348378	25.153079
20	−7.348372	25.144737

If coefficients s_D* are regarded as independent sources, then a procreative linear statistical model is acquired. Additionally, if A_m is assumed to be squared and invertible, the standard ICA model [1] is required. In this work, fast ICA approach [17, 18] is applied to the disintegrated signals to calculate mixing and separating matrices (A_m and *X,* respectively) as well as to the matrix of independent components. Subsequently, the significant sources are selected according to obtained *SNR* values. Hence, the independent component with the maximum *SNR* value is the reconstructed signal.

3 Results and Discussions

The EEG recordings during visual relaxation used in this work were downloaded from PhysioBank ATM [19, 20]. The database comprises EEG recordings from 14 subjects with duration of 10 s each. The signals were exported in the *.csv* format. Using the database, different magnitudes of AWGN noises with *SNR* values ranging from 5 to 20 dB were added by simulation. Decomposition using SWT was performed with mother wavelet *symlet* and decomposition level 6. Soft thresholding was then applied to the decomposed EEG signals. Fast ICA using the kurtosis technique was adopted and implemented to attain maximized non-Gaussianity. The convergence criteria ($C = 0.00$ and max. iterations = 100) were assumed to minimize Gaussianity. After this, the reconstructed EEG signals are obtained as presented in Fig. 2.

The performance evaluation and comparison of various obtained results are done using the parameter *SNR*. Figure 2 shows experimentation of *SNR*(noise) = 10 dB on two different signals. The *SNR* of reconstructed signal was 25.20 and 26.05 dB, respectively. This experiment was repeated for various values of *SNR*, i.e., 5, 10, 15, and 20 dB, respectively. Table 1 shows the comparison of results on the basis of several *SNR* values. The observed trend shows that on increasing the *SNR*(noise), the *SNR* of reconstructed signal goes on decreasing. These obtained results clearly present the efficiency of the proposed approach. Higher *SNR* values of the reconstructed signals prove that the applied algorithms have worked efficiently.

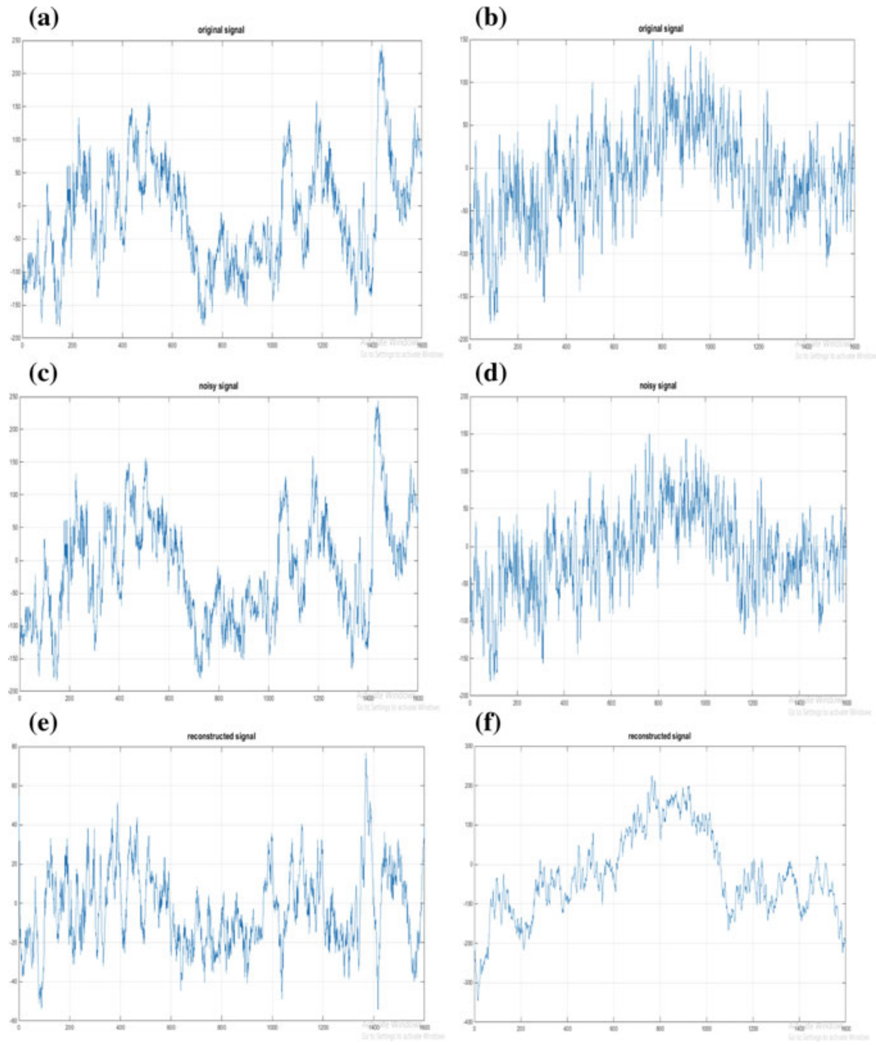


Fig. 2 **a**, **b** Original EEG signals 1 and 2, respectively of *SNR* = 10 dB. **c**, **d** Noisy EEG signals 1 and 2, respectively, of *SNR* = 10 dB. **e**, **f** Reconstructed EEG signals 1 and 2, respectively, of *SNR* = 10 dB

4 Conclusion

In this work, a methodology has been proposed for the denoising of single-channel EEG signal on the principle of combined usage of SWT–ICA approach. Here, the artifacts modeled as Gaussian noises on varying *SNR* values are added to the original EEG signal and then processed through the proposed algorithm. This approach presents high *SNR* values of the reconstructed EEG signals over the single-channel small-scale database. The maximum value of *SNR* of reconstructed signal is obtained at the lowest value of *SNR*(noise), i.e., 5 dB. The conclusion can be made that the proposed approach eliminates the noise at higher rate on lower *SNR* values. For future prospects, the processed EEG signal can be used for numerous applications like human biometric authentication, clinical and research applications, etc. Proper preprocessing enhances recognition accuracy for biometric applications. The proposed approach facilitates the usage of single-channel EEG for various applications.

References

1. Zahhad, M.A., Ahmed, S.M., Abbas, S.N.: A new multi-level approach to EEG based human authentication using eye blinking. A J. Pattern Recogn. Lett., **82**, 216–225 (11 Aug 2015)
2. Sheoran, M., Kumar, S., Kumar, A.: Wavelet-ICA based denoising of electroencephalogram signal. Int. J. Inf. Comput. Technol. **4**(12), 1205–1210 (2014)
3. Lay-Ekuakille, A., Vergallo, P., Griffo, G., Urooj, S., Bhateja, V., Conversano, F., Casciaro, S., Trabacca, A.: Multidimensional analysis of EEG features using advanced spectral estimates for diagnosis accuracy. In: Proceedings of Medical Measurements and Applications, pp. 237–240. IEEE (2013)
4. Bhateja, V., Verma, R., Mehrotra, R., Urooj, S.: A non-linear approach to ECG signal processing using morphological filters. Int. J. Measur. Technol. Instrum. Eng. **3**(3), 46–59 (2013)
5. James, C. J., Gibson, O. J.: Temporally constrained ICA: an application to artifact rejection in electromagnetic brain signal analysis. IEEE Trans. Biomed. Eng. **50**(9), 1108–1116 (2003)
6. Klass, D.W.: The continuing challenge of artifacts in the EEG. Am. J. EEG Technol. **35**(4), 239–269 (1995)
7. Islam, M.K., Rastegarnia, A., Yang, Z.: A wavelet-based artifact reduction from scalp EEG for epileptic seizure detection. IEEE J. Biomed. Health Inform. **10**(5), 2168–2194 (2015)
8. Hyvarinen, A., Oja, E., Parkkonen, L., Hari, R.: Independent component analysis: algorithms and applications neural networks. A J. Neuroimage **13**(4–5), 411–430 (2000)
9. Devuyst, S., Dutoit, T., Stenuit, P., Kerkhofs, M., Stanus, E.: Cancelling ECG artifacts in EEG using a modified independent component analysis approach. EURASIP J. Adv. Sig. Process. **2008**(1), 1–13 (2008)
10. Verma, R., Mehrotra, R., Bhateja, V.: An improved algorithm for noise suppression and baseline correction of ECG signals. In: Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA), pp. 733–739, Springer, Berlin, Heidelberg (2013)
11. Verma, R., Mehrotra, R., Bhateja, V.: An integration of improved median and morphological filtering techniques for electrocardiogram signal processing. In: 3rd International Advance Computing Conference (IACC), pp. 1223–1228, IEEE (2013)
12. Hyvarinen, A., Ramkumar, P., Parkkonen, L., Hari, R.: Independent component analysis of short time fourier transform for spontaneous EEG/MEG analysis. A J. Neuroimage **49**(5), 257–271 (2010)
13. Anand, D., Bhateja, V., Srivastava, A., Tiwari, D.K.: An approach for the preprocessing of EMG signals using canonical correlation analysis. In: Smart Computing and Informatics, pp. 201–208, Springer, Singapore (2018)
14. Wang, X.H., Istepanian, R.S.H.: Microarray image enhancement by denoising using stationary wavelet transform. IEEE Trans. Nanosci. **2**(4), 184–189 (2003)
15. Lay-Ekuakille, A., Vergallo, P., Griffo, G., Conversano, F., Casciaro, S., Urooj, S., Bhateja, V., Trabacca, A.: Entropy index in quantitative EEG measurement for diagnosis accuracy. IEEE Trans. Instrum. Meas. **63**(6), 1440–1450 (2014)
16. Donoho, D.L.: Denoising by soft-thresholding. IEEE Trans. Inf. Theory **42**(3), 613–617 (1995)

17. Inuso, G., Mammone, N., Morabito, F.C., Foresta, F.L.: Wavelet-ICA methodology for efficient artifact removal from electroencephalographic recordings. In: Proceedings of International Joint Conference on Neural Networks, pp. 1524–1529 (2007)
18. Kumar, P.S., Arumuganathan, R., Sivakumar, K., Vimal, C.: A wavelet based statistical method for DeNoising of ocular artifacts in EEG signals. Int. J. Comput. Sci. Netw. Secur. **8**(9), 87–92 (2008)
19. Lay-Ekuakille, A., Griffo, G., Conversano, F., Casciaro, S., Massaro, A., Bhateja, V., Spano, F.: EEG signal processing and acquisition for detecting abnormalities via bio-implantable devices. In: International Symposium on Medical Measurements and Applications (MeMeA), pp. 1–5, IEEE (2016)
20. PhysioBank ATM, https://physionet.org/cgi-bin/atm/ATM

Non-local Mean Filter for Suppression of Speckle Noise in Ultrasound Images

Avantika Srivastava, Vikrant Bhateja, Ananya Gupta and Aditi Gupta

Abstract The speckle suppression is very important for carrying out proper clinical operation in ultrasound images. Speckle is a kind of multiplicative noise which exists inherently within the ultrasound images. Although there are many speckle noise suppression filters from the ultrasound images, it possesses certain constraints. In this paper, the Non-local Mean (NLM) filter is used to suppress speckle noise from ultrasound images. NLM filter consists of two windows that are search and similarity windows which can improve upon the limitation of many conventional filters. The simulation results of NLM filter for the values of different noise variances on ultrasound image are shown. Finally, the performance analysis is done by image quality assessment parameters like Peak Signal-to-Noise Ratio (PSNR) and Coefficient of Correlation (CoC).

1 Introduction

Ultrasound is a diagnostic imaging technique which works on application of ultrasonic waves. Ultrasound is a technique used to see the internal body structures, for instance, muscles, tendons, vessels, joints, and internal organs. Image acquisition is conducted using the ultrasound equipment which includes transducer, scanner, CPU, and display device. In the acquisition process, CPU sends current to the transducer probe and also collects the electrical pulses from the probes. Ultrasound waves are

A. Srivastava · V. Bhateja (✉) · A. Gupta · A. Gupta
Department of Electronics and Communication Engineering, Shri Ramswaroop Memorial Group of Professional Colleges (SRMGPC), Lucknow 226028, Uttar Pradesh, India
e-mail: bhateja.vikrant@gmail.com

A. Srivastava
e-mail: avantika.srivastava1996@gmail.com

A. Gupta
e-mail: ananyag530@gmail.com

A. Gupta
e-mail: agaditi12@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_23

produced from the transducer and travel through body tissues, and when the waves reach an object or surface with different textures, it is reflected back. These echoes are received by the apparatus, i.e., the transducer arrays, and changed into current signals which are shown on the display device. When the ultrasound waves travel through tissues, some of the waves are partly transmitted to the deeper structure and some waves are reflected back to the transducer as echoes which will produce the ultrasound image [1]. A small portion of the returning sound pulse by transducer surface gets reflected back into the tissues and generates a new echo at twice the depth, which is called as speckle noise. The nature of speckle noise is multiplicative, which perceptually appears as variation in contrast of the image. Images which contain speckle noise will result in reducing the contrast of image, and hence it becomes difficult to perform image processing operations like edge detection, segmentation, feature extraction, etc. Speckle noise holds a granular pattern which is the intrinsic property of ultrasound image. Implementing a speckle suppression filter can reduce the deprivation of ultrasound image by improving the contrast and thus retaining the fine details of the image. Many post-processing algorithms of ultrasound image such as image segmentation, registration, or classification of tissue parenchyma are totally based on real ultrasound images, and these algorithms can get affected by the presence of the speckle noise. Therefore, for proper clinical analysis and quantitative measurements, it is important to suppress speckle noise from the ultrasound images [2, 3]. The speckle suppression filters can be broadly characterized as local statistics filters which are mean [4], median [5], Lee [6], Kuan [7, 8], Frost [8], Wiener [9], LSMV [7], and anisotropic diffusion filters which are anisotropic diffusion [10] and SRAD [11]. The major drawback of local statistic filter is that they cannot differentiate between edges and noise present in ultrasound image. Also, over-smoothing of ultrasound images leads to blurred or dull images. The anisotropic diffusion filters can differentiate between edges and noise by the use of diffusion coefficient. But the challenge observed is that if iterations are increased then image gets blurred, and hence the computation complexity increases. That is why, another category of filter was introduced in the last decade, called as NLM filter. The NLM filter can reduce these drawbacks. The NLM filter was introduced by Antony Buades in [12], in which a different non-local approach is developed for noise suppression. This technique was totally based on similarity of pixels in the image having an advantage to suppress noise. In this, each window consists of similar window which improves the inadequacy of many traditional filters. The local statistics filter and anisotropic diffusion filter are not capable for effective suppression of speckle noise. So, the NLM filter is the remedial solution for the speckle suppression in ultrasound images. In this paper, the NLM filter is applied to the ultrasound images for the suppression of speckle noise. The results are obtained for different levels of noise variance in ultrasound images. Finally, the performance of NLM filter is analyzed using image quality assessment parameters like Peak Signal-to-Noise Ratio (*PSNR*) and Coefficient of Correlation (*CoC*). The further segments of this paper are as follows: in the second segment, the explanation of NLM filter is discussed which also includes the algorithm of NLM filter. In the third segment, various results are obtained on

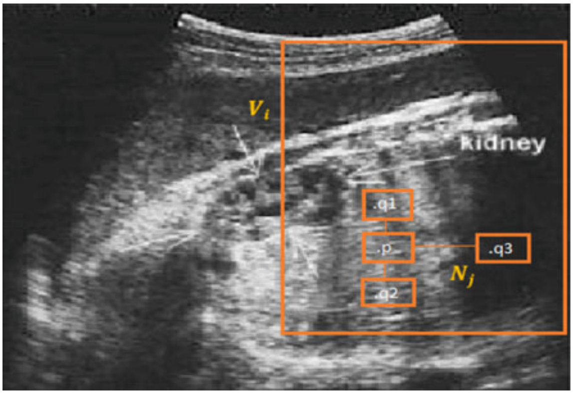


Fig. 1 NLM filter concept [12]

different noise variances, and the performance of the filter is analyzed and the last segment is conclusion.

2 NLM Filter for Speckle Noise Suppression

NLM is basically a filtering algorithm for image denoising whose concept is entirely based on the self-similarity. NLM filter takes only those particular pixels which are similar to the target pixel according to their geometrical configuration and pixel intensity. Contrasting from the other local mean filter which takes the mean average of surrounding pixels around the target pixel, NLM filter is more capable of saving fine details of the image accompanied by enhancing the contrast of the real image in comparison with other local mean filter. NLM filter is totally based on the similarity of the pixels in the images taking a high degree of advantage of the redundancy to suppress noise. Here, every window consists of similar window which can improve upon the limitations of many conventional filters. The NLM filter tends to compare the intensity level and the geometrical structure of the entire pixel neighborhood. The restored pixel intensities which are obtained are the average weights of all the pixel intensities of the original image. NLM filter thus brings the noise suppression with detail preservation. Here, an example of NLM filter is shown in Fig. 1. There are two windows in the image, larger one is called search window and smaller window is called similarity window. Similarity window moves through the search window and compares each pixel to the target pixel. Here, four pixels are being given, i.e., $q1$, $q2$, $q3$, and p, respectively. The neighbor pixels $q1$ and $q2$ are similar to that of target pixel p but not $q3$ as it differs in pixel intensity and geometrical configuration. The mean average of the similar pixels is been obtained and replaced by the target pixel, and the filtered image can be obtained with less loss of details [12, 13].

The algorithm which is involved in the process of NLM filter has been discussed [13–15]. For a noisy image Y, for pixel i denoised value $Y(i)$ can be achieved by the following equations. Computation of pixel weight can be calculated by the formula

$$w(i, j) = \frac{1}{Z(i)} \mathrm{e}^{-\frac{\left|Y(N_i)-Y(N_j)\right|_{2,\sigma}^{2}}{h^2}} \tag{1}$$

Here, $w(i,j)$ denotes calculated weight of neighborhood pixel i and j, respectively, $\left|Y(N_i) - Y\left(N_j\right)\right|_2^2$, σ is the Euclidean distance and normalizing constant is denoted by $Z(i)$ which is calculated by the following formula:

$$Z(i) = \sum \mathrm{e}^{-\frac{\left|Y(N_i)-Y\left(N_j\right)\right|_{2,\sigma}^{2}}{h^2}} \tag{2}$$

After calculating the normalized weight, denoised value $Y(i)$ for pixel i has been calculated by the following formula:

$$\mathrm{NLM}(Y(i)) = \sum_{j \in Y} w(i, j) Y(j) \tag{3}$$

3 Results and Discussions

The NLM filtering involves certain parameters which are fixed such as window size of search and similarity window, filtering parameter h. The search window size is fixed to 3, and size of similarity window is fixed to 2. The filtering parameter h is taken as 10. Original ultrasound image which is taken is of fetal kidney [16]. The various results on the ultrasound image are shown in Table 1. The NLM filter is applied to the ultrasound image with different noise levels of variance (σ) which ranges from $\sigma = 0.001$, $\sigma = 0.02$, $\sigma = 0.04$, $\sigma = 0.06$, and $\sigma = 0.08$. The value of $\sigma = 0.001$ implies the very low amount of noise present in the ultrasound image. Figure 2b, c shows the result for low variance of noise in ultrasound image. Here, σ value is increasing up to the maximum level $\sigma = 0.08$, and the result for the maximum variance of noise is shown in Fig. 2j, k. It can be seen here that as the value of σ is increasing, the performance of the NLM filter is declining although at the same time the performance of NLM filter for low σ is good.

Table 1 Algorithm of NLM filter

BEGIN	
Step 1: *Input* Noisy Ultrasound Image (Y) with multiplicative noise. **Step 2:** *Convert* Image to Gray Scale Image (Y’). **Step 3:** *Initialize* Search Window (t) and Similarity Window (f). **Step 4:** *Initiate* Patches for search window and Patches for similarity window. **Step 5:** *Compute* of Pixel Weight given by Eq. (1). **Step 6:** *Compute* of Normalized Weight given by Eq. (2). **Step 7:** *Compute* of Weighted Average given by Eq. (3). **Step 8:** *Check* whether we have reached last pixel? If yes then go to step 9, else again start from step 4. **Step 9:** *Display* Denoised Ultrasound Image (Y).	
END	

Figure 2 depicts the performance of NLM filter at different noise variances on ultrasound image. It is known that, as the value of PSNR increases, the quality of image will also increase. It is shown in Table 1 that the value of PSNR is high for the image with lowest amount of noise and as the amount of noise in image is increasing the value of PSNR is decreasing. Hence, the conclusion is made that the NLM filter is better for the lower amount of noise as it cannot suppress the noise effectively of higher variance of noise. The CoC shows the relationship among the real value of image and the actual value of image. It is shown in Table 1 that the value of CoC is decreasing as the amount of noise present in the image is increasing.

4 Conclusion

In this paper, simulation of NLM filter is performed by adding different values of noise variance to the original ultrasound image. The results are being obtained by calculating image assessment parameters such as PSNR and CoC as shown in Table 2. The aim of the paper is to develop NLM filter for speckle noise suppression which should be capable of retaining the fine structure of the original image without losing

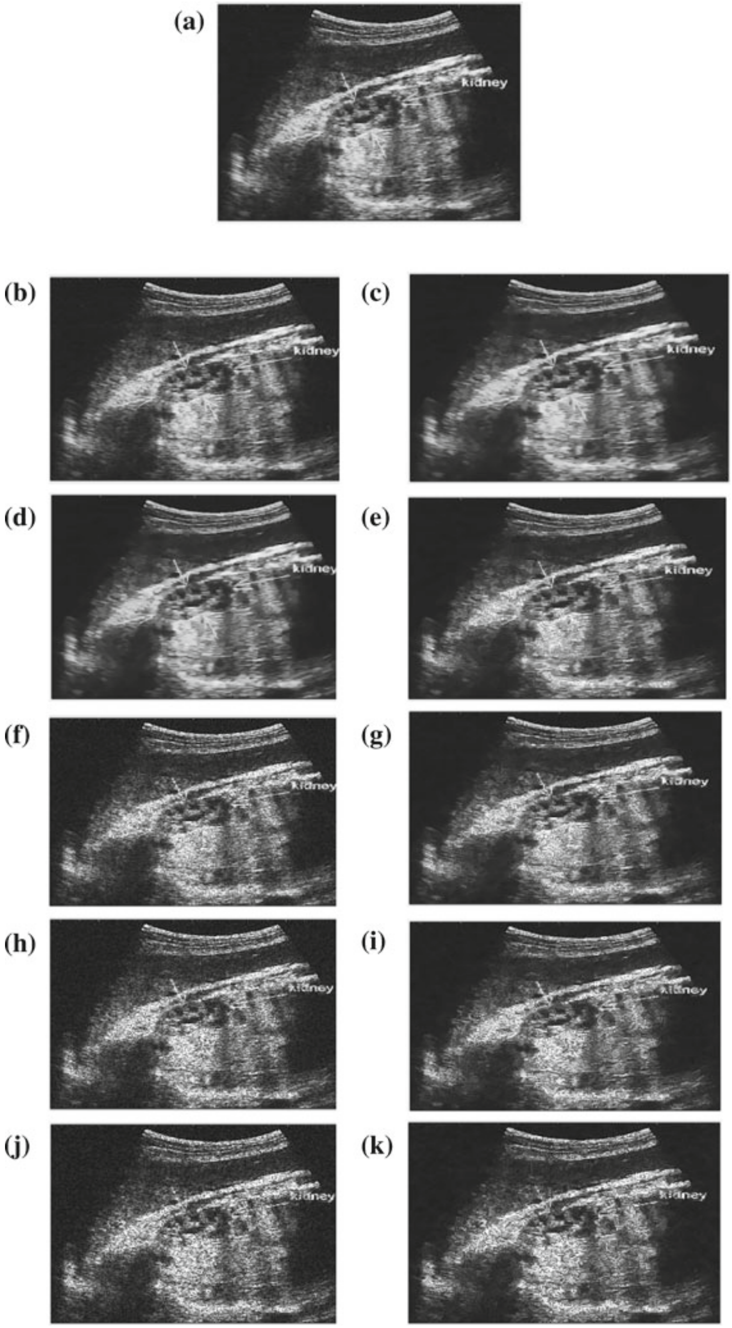


Fig. 2 **a** Original ultrasound image, noisy images with, **b** noise variance $= 0.001$, **d** noise variance $= 0.02$, **f** noise variance $= 0.04$, **h** noise variance $= 0.06$, **j** noise variance $= 0.08$; filtered images (**c**), (**e**), (**g**), (**i**), and (**k**)

Table 2 Simulation result of PSNR and CoC for ultrasound image

Noise variance	PSNR (in dB)	CoC
0.001	58.1197	0.9909
0.02	48.8228	0.9715
0.04	46.8574	0.9475
0.06	44.6197	0.9238
0.08	42.9808	0.9011

any fine details of the image along with the restoration of the edges and contrast enhancement of the image. For the future aspect of this paper, NLM will perform better and work effectively on the images with high variance noise.

References

1. Bhateja, V., Tripathi, A., Gupta, A., Lay-Ekuakille, A.: Speckle Suppression in SAR Images Employing Modified Anisotropic Diffusion Filtering in Wavelet Domain for Environment Monitoring, in Measurement, New Delhi, India, vol. 74, pp. 246–254 (2015)
2. Bhateja, V., Singh, G., Srivastava, A., Singh, J.: Speckle reduction in ultrasound images using an improved conductance function based on anisotropic diffusion. In: Proceedings of International Conference of Computing for Sustainable Global Development (INDIACom), pp. 619–624. IEEE (2014)
3. Bhateja, V., Tiwari, H., Srivastava, A.: A non local means filtering algorithm for restoration of Rician distributed MRI. In: Emerging ICT for Bridging the Future-Proceeding of the 49th Annual Convention of the Computer Society of India CSI, vol. 2, pp. 1–8. Springer, Cham (2015)
4. Zhang, P., Li, F.: A new adaptive weighted mean filter for removing salt and pepper noise. IEEE J. Sig. Process. Lett. **21**(10), 1280–1283 (2014)
5. Loupas, T., McDicken, N.W., Allan, L.P.: An adaptive weighted median filter for speckle suppression in medical ultrasound images. IEEE Trans. Circ. Syst. **36**(1), 129–135 (1989)
6. Lee, S.J.: Digital image enhancement and noise filtering by use of local statistics. IEEE Trans. Pattern Anal. Mach. Intell. **2**(2), 165–168 (1980)
7. Loizou, P.C., Pattichis, S.C., Christodoulou, I.C., Istepanian, S.R., Pantziaris, M., Nicolaides, A.: Comparative evaluation of despeckle filtering in ultrasound imaging of the Cartoid artery. IEEE Trans. Ultrason. Ferroelectr. Freq. **52**(10), 1653–1669 (2005)
8. Finn, S., Glavin, M., Jones, E.: Echocardiographic speckle reduction comparison. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **58**(1), 82–101 (2011)
9. Sivakumar, J., Thangavel, K., Saravanan, P.: Computed radiography skull image enhancement using Wiener filter. In: Proceedings of International Conference on Pattern Recognition, Informatics and Medical Engineering, Tamil Nadu, India, pp. 307–311. IEEE (2012)
10. Tripathi, A., Bhateja, V., Sharma, A.: Kuan modified anisotropic diffusion approach for speckle filtering. In: Proceedings of the First International Conference on Intelligent Computing and Communication, pp. 537—545. Springer, Singapore (2017)
11. Bhateja, V., Sharma, A., Tripathi, A., Satapathy, S.C., Le, D.N.: An optimized anisotropic diffusion approach for despeckling of SAR images, In: Digital Connectivity Social Impact, pp. 134–140. Springer, Singapore (2016)
12. Baudes, A., Coll, B., Morel, M.J.: A non-local algorithm for image denoising. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 60–65. IEEE (2005)

13. Baudes, A., Coll, B., Morel, M.J.: A review of image denoising algorithms, with a new one. SIAM J. Multiscale Model. Simul.: A SIAM Interdisc. J. **4**(2), 490–530 (2005)
14. Bhateja, V., Mishra, M., Urooj, S., Lay-Ekuakille, A.: Bilateral despekling filter in homogeneity domain for breast ultrasound images. In: Proceedings of International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1027–1032. IEEE (2014)
15. Bhateja, V., Sharma, A., Tripathi, A., Sapathy, S.C.: Modified non linear diffusion approach for multiplicative noise. In: Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications, pp. 343–350. Springer, Singapore (2017)
16. Ultrasound Pictures, www.ob-ultrasound.net/frames.htm

Various Image Segmentation Algorithms: A Survey

Kurumalla Suresh and Peri Srinivasa rao

Abstract Image segmentation is a necessary method in image processing. It is nothing but partitioned an image into several parts called segments. It has applications like image compression; because of this type of application, it is unable to develop the entire image. In that, time segmentation technique is used, to segment the portions from the image for remaining processing. Already certain methods are existed, which divides the single image into multiple parts depending on some constraints like intensity value of the pixel, image color, size, texture, etc. These methods can be divided based on segmentation method. In this paper, author reviewed some algorithms, and finally their pros and cons are listed.

1 Introduction

An image is a mode of conveying details, and the image holds lots of effective information. Understanding the image and obtaining data from the image to achieve certain works is a principal area of application in digital image technology, and the main step in interpreting the image is called as image segmentation. Nowadays, image segmentation is popular and challenging field in image processing. Practically, it is often not interested in all parts of the image, but only for some certain areas which have the same characteristics. It is foundation for image processing. It is also an important basis for image recognition. It depends on certain measure to partition an input image into several numbers of the equal nature of group in order to bring out the area in which people are focused on. It is the foundation for image analysis and understanding of image feature extraction and recognition. The concept of image

K. Suresh (✉)
Jawaharlal Nehru Technological University, Kakinada, A.P, India
e-mail: kurumallasuresh@gmail.com

P. Srinivasa rao
CS&SE dept, Andhra University College of Engineering, Visakhapatnam, A.P, India
e-mail: peri.srinivasarao@yahoo.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_24

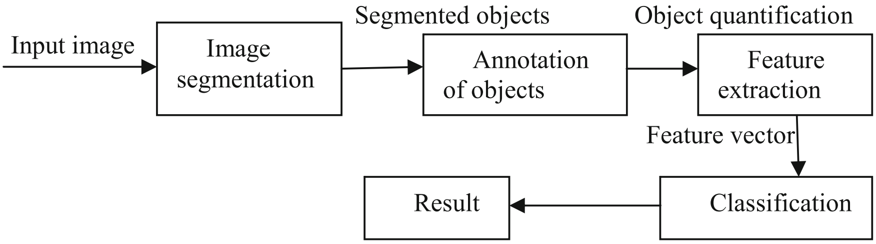


Fig. 1 Image analysis pipeline

segmentation is simplification. It is the first and main step in image analysis and processing (Fig. 1).

Every algorithm in image segmentation has three characteristics. They are correctness, stability toward parameter choice, and stability toward image choice.

Correctness
Segmentation clearly finds the structure in the image. It is not too good or bad up to certain level of details.

Stability toward parameter choice
The aim is to generate a segmentation of uniform definiteness for a wide area of parameter choice.

Stability toward Image choice
The aim is to generate a segmentation of uniform correctness in a wide range of various images.
If any algorithm obeys these three characteristics, then only it will display the efficient expected result, which is incorporated into larger systems.
Image segmentation algorithms have two basic properties. They are discontinuity and similarity. Discontinuity is nothing but portioning an image based on edge intensity values, whereas similarity is nothing but portioning an image based on region methods like region growing and region splitting and merging.

2 Image Segmentation Techniques

Partitioning a single image into several parts is called image segmentation. It is based on multiple methods. They are threshold, edge-based, region-based, clustering-based, watershed-based, PDE-based, and ANN-based methods.

Threshold method
Thresholding method [1] is easy and efficient procedure for image segmentation. In this, image pixels are partitioned with the support of image intensity. This technique is mainly focused on how to separate front objects from background. It can be divided into three types. They are

(1) **Global thresholding**:
It completely depends upon threshold value selection.

(i) If pixel value > threshold, then it indicates one.
(ii) If pixel value ≤ threshold, then the output indicates zero.

(2) **Variable thresholding**:
In global thresholding, T value is fixed, but in this it varies through an image.

(3) **Multiple thresholding**:
As the name itself, it indicates many threshold values from t_0 to t_n.

Region-based segmentation
In this method, grouping the smaller regions into big regions is done.
It can be divided into two parts [2].

(i) Seeded region growing method and
(ii) Unseeded region growing method.

In seeded region growing process, first start with a set of seeds. Then, the region is grown by adding remaining seeds which are having similar properties. After that, perform splitting and merging process.

In unseeded region growing process, no need to select initial seeds for segmentation. Remaining process is same as seeded region growing method.

Clustering-based image segmentation
Clustering is another synonym for image segmentation. In this method [3], group the objects which are having similar characteristics. The techniques used for clustering same techniques are applicable for image segmentation.

Edge-based image segmentation
These are famous and advanced methods in image processing. It depends upon intensity change in the image. Single intensity value does not give the efficient result about image. In this, first we find the edges after connecting the edges with object boundaries to segment the image. Two methods are used here; they are gray histograms and gradient-based methods. The output is displayed as binary image. These techniques depend upon the discontinuity detection [4].

Watershed-based method
The watershed method [5] depends on topological interpretation approach. In this algorithm, intensity indicates with the basin containing holes where water comes into the outside. When the water reaches to basin boundary, then the adjacent regions are combined. These methods used the image gradients like topographic surface.

Partial Differential Equation Based Segmentation method
These methods are speed techniques of image segmentation. These are mainly used for crucial time applications. These are mainly divided into two types. They are (1) nonlinear isotropic diffusion filter and (2) convex non-quadratic variation restoration (It eliminates the noise). Then, the output displays a blurred image. The fourth-order

PDE is used mainly for decreasing noise of the image. Second-order PDE mainly identifies the edges [5].

Artificial Neural Network Based Segmentation method: [2]
The main aim of this method is decision-making. It is mainly used in medical fields to remove the background from the required image. It is independent of the partial differential equation based segmentation method.
There are regularly used image segmentation algorithms. This paper relates the following four algorithms for simple survey.

3 Related Work

Zhensong Chen et al.:

Zhensong Chen et al. developed an algorithm for image segmentation. It depends upon the DP clustering algorithm [6]. In this algorithm, there is no need of prior knowledge about cluster numbers. Only two parameters are considered for each point i. One is density ρ_i and the second one is distance δ_i.

$$\text{Density}\rho_i = \Sigma_j \exp - \frac{d_{ij}^2}{d_c^2} \tag{1}$$

d_{ij} indicates distance between i and j,
d_c indicates cutoff distance, and
ρ_i indicates points distribution around i.
distance δ_i can be calculated using the formulae

$$\delta_i = \begin{cases} \min_j(d_{ij}) & \dfrac{\rho_j > \rho_i}{\rho_i \text{ is the higher density}} \\ \max_j(d_{ij}) & \end{cases} \tag{2}$$

Algorithm steps:

(1) In this algorithm, first consider input image data to gain the indications in three color channels.
(2) Find the distance and density using Eqs. (1) and (2). Then, compute the decision graph.
(3) Select the points with high density and large distance. And then find the cluster number.
(4) If $\rho_j > \rho_i$ and ρ_i has higher density, these two conditions are satisfied at the point x_i. Then, the points are assigned to the same label x_j.
(5) Finally, segmentation depends on the label marks.

Digabel et al.: [7]
Digabel et al. proposed region-based image segmentation algorithm. It completely

depends upon the watershed geographical concept [8, 9]. In this first, figure out the segmentation function of the image. It is nothing but dark regions are considered as objects in the image. Then, identify foreground and background marks. Foreground markers are pixels which are part of any object. At the end, figure out the watershed transformation.

Suresh et al.: [10]
Suresh et al. proposed a new algorithm called efficient DBSCAN for image clustering. It depends upon the traditional DBSCAN approach. In this method, first consider an RGB image and then converted into the gray color image. If noise is presented in the gray image, remove the noise. Then, calculate the minpts and eps. Minpts are based on image size.

Let image size is $M * N$

Then, minpts are calculated using the formulae

$$\text{Minpts} = \frac{M * N}{256} \tag{3}$$

256 indicates the gray level value of the image, and $M * N$ indicates the pixel image size. Eps depends upon minpts and KNN algorithm. Then, apply the traditional DBSCAN approach [11, 12]. Using this clustering technique, the image is segmented and displayed as output.

Improved Fuzzy C-Means algorithm [13]
The famous clustering techniques like FCM [14, 15] and FLICM [16] are not suitable for segmentation with noise images. In both the algorithms, image gray values are taken into consideration. So IFCMA is introduced. It gives the efficient results with noise images also. For this, Euclidean distance measure is used to find the distance between pixels. In this algorithm, first compute the total number of clusters. Then, compute the center of cluster with the help of Euclidian distance and find the membership matrix. The values which are present in the matrix update the center of cluster with that values. Repeat the algorithm until Euclidian distance is greater than the matrix value.

Comparative Analysis of algorithms with advantage and Disadvantages
Table 1 shows the advantages and disadvantages of different image segmentation algorithms.

4 Research Challenges in Image Segmentation

After reviewing the above image segmentation algorithms, authors are identified some research challenges in this area. They are as follows:

Scalability
Majority of the image segmentation algorithms work effectively for small size images

Table 1 Comparative analysis of image segmentation algorithms

Algorithm	Methodology	Advantage	Disadvantage
Zhensong Chen et al.	Distance and density	• No need of prior knowledge about cluster number	• High time complexity • Less efficiency
Digabel et al.	Watershed geographical concept	• Used to separate foreground and background objects	• Need to specify number of clusters for segmentation function
Suresh et al.	Eliminating noise points	• Noise resistant	• It is expensive • It takes more time
Improved fuzzy C-means algorithm	Euclidean distance measure	• Effective results with noise images also	• Usage of membership matrix

only, but in real scenario images are vast in size. Hence, present algorithms are less effective to segment the image and even few algorithms unable to segment the images also. Hence, there is a scope for developing scalable algorithms to segment the images in present real world.

Quality
The proposed image segmentation algorithms are usually tested on well-known pixels of image or image segmentation with very small structure but those algorithms not tested the quality on large images. Hence, there is a scope for testing the algorithm with different benchmarks to reveal the quality of images in real-world networks.

5 Conclusion and Future Scope

In this survey, authors presented different image segmentation algorithms on small size images only. Majority of present algorithms works well on small size images but not suitable for large images. These image segmentation techniques are based on distance and density, watershed geographical concept, eliminating noise points, and Euclidean distance. The spatiality of this work is that it reveals the literature review of different image segmentation algorithms and provides a large amount of information under a single paper. After reviewing all the existing algorithms, this survey concludes that scalability and efficiency are the major factors affecting community detection.

References

1. Zhang, Y.J.: An overview of video and Image segmentation in the last 40 years. In: Proceedings of the 6th International Symposium Applications on Signal Processing, pp. 1441–1451 (2001)
2. Rajesh, R., Senthilkumaran, N.: Edge detection techniques for image segmentation- a survey of soft computing approaches. Int. J. Recent Trends Eng. **1**(2) (2009)
3. Yambal, M., Guptha, H.: Image segmentation using FCM clustering: a survey. Int. J. Adv. Res. Comput. Commun. Eng. **2**(7) (2013)
4. Kalyankar, N.V., Saleh, S., Khamitkar, S.: Image segmentation using edge detection (UCSE). Int. J. Comput. Sci. Eng. **02**(03) (2010)
5. Liang, R.R., Yang, Q.Q.: The comparative research on image segmentation algorithms. In: IEEE Conference on ETCS 2009, pp. 703–770
6. Dilpreet Kaur et al.: Various image segmentation techniques: a review. IJCSMC **3**(5), pp. 809–814 (2014)
7. Hui, S.M., Bhandarkar, Z.: Image segmentation using evolutionary computation. IEEE Trans. Evol. Comput. **3**(1), pp. 1–21 (1999)
8. Wang, D.: A multiscale gradient algorithm for image segmentation using watersheds. Patt. Recogni. **30**(12), 2043–2052 (1997)
9. Kim, J.B., Kim, H.J.: Multi resolution based watersheds for efficient image segmentation. Patt. Recogni. Lett. **24**, 473–488 (2003)
10. Kurumalla, S., et al.: K-nearest neighbor based DBSCAN clustering algorithm for image segmentation. J. Theor. Appl. Inf. Technol. **92**(2) (2016)
11. Mohamedt, R.F., ElBarawy, Y.M., Ghali, N.I.: Improving social network community detection using DBSCAN algorithm. In: 2014 World Symposium Computer Applications & Research (WSCAR). IEEE (2014)
12. Mehjabin., et al.: Community detection methods in social networks. I.J. Educ. Manag. Eng. **1**, 8–18 (2015). (http://www.mecs-press.net), https://doi.org/10.5815/ijeme
13. Li, L., Hu, X.: Improved F c-means algorithm for image segmentation. J. Electr. Electron. Eng. **3**(1), 1–5 (2015)
14. Dunn, J.C.: A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters. J. Cybern. **3**, 32–57 (1973)
15. Bezdek, J.C.: Pattern Recognition with F Objective Function Algorithms. Kluwer Academic Publishers (1981)
16. Chatzis, V., Krinids, S.: A robust of logical information C-means clustering algorithm. IEEE Trans. Image Process. **19**(5), 1328–1337 (2010)

Peak-to-Average Power Ratio Performance of Transform Precoding-Based Orthogonal Frequency Division Multiplexing Offset Quadrature Amplitude Modulation

N. Renuka and M. Satya Sairam

Abstract The noncontiguous orthogonal frequency division multiplexing (NC-OFDM) emerged as an excellent technique in digital communications. This technique often uses offset quadrature amplitude modulation (NC-OFDM/OQAM). This is technique is considered as the effective multicarrier method implemented in cognitive radio communications. Similarly, NC-OFDM/OQAM also suffers due to peak-to-average power ratio (PAPR). In this article, by employing two different transform techniques, the PAPR is reduced. Here, overlapped selective mapping (OSLM) technique is combined with Walsh–Hadamard transform (*WHT*) and Zadoff–Chu transform (ZCT) which are proposed to control the PAPR in NC-OFDM/OQAM systems. The proposed method combines OSLM with all transforms in different ways. In this method before OSLM, the transforms *WHT* and ZCT are applied. A comparative study of these simulation results has been performed to elevate the performance of the proposed characteristics.

1 Introduction

The OFDM aided with OQAM posses interference-free bandwidth characteristics [1]. Along with this, it is also featured with very low sidelobes [2]. Also, it reports excellent efficiency [3]. Wireless systems employing these techniques are very much suitable for cognitive radio (CR) applications. It is noticeable that the signal form differs from conventional modulation scheme [4]. In spite of these features, it is very much prone to PAPR-based issues. In the literature, several methods are proposed to control this PAPR [5–9]. However, this technique cannot be embedded in the

N. Renuka
A.N.U. College of Engineering & Technology, Acharya Nagarjuna University, Guntur, A.P, India
e-mail: renukanellaturu@gmail.com

M. Satya Sairam (✉)
Chalapathi Institute of Engineering and Technology, Guntur, India
e-mail: msatyasairam1981@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_25

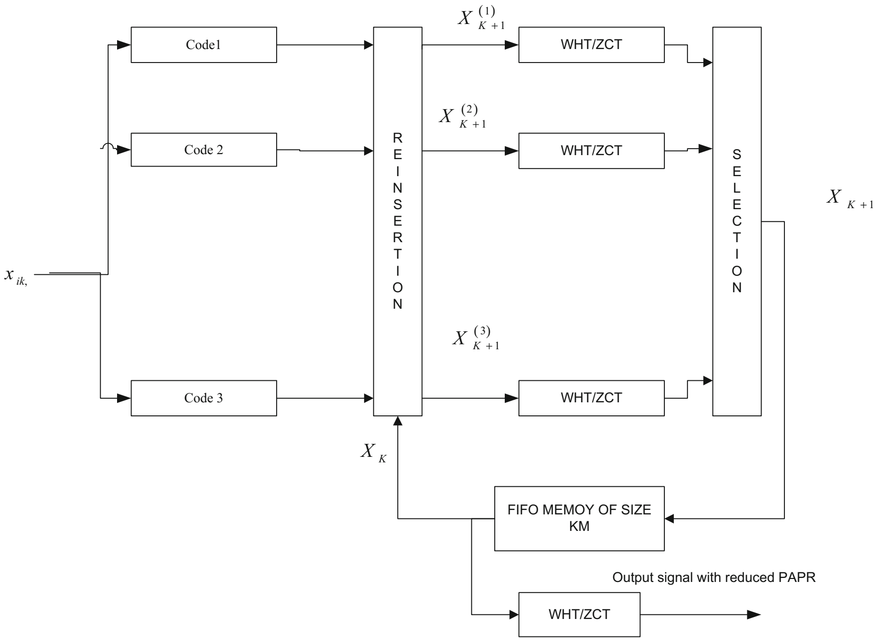


Fig. 1 WHT and ZCT overlapped selective mapping technique

technique in a simple manner. The possible consequences are like clipping, distortion, etc.

In this paper, two novel transformation techniques are applied to control the PAPR in the considered version of OFDM. Further, the paper is organised as follows. Brief description of the PAPR in NC-OFDM/OQAM is given in Sect. 2, and the mapping phenomenon is given in Sect. 3. Simulation results and overall conclusions are given in Sects. 4 and 5, respectively.

2 PAPR of NC–OFDM/OQAM

The OQAM/OFDM has drawn significant attention recently with the obvious advantages discussed above in the previous section. Owing to the overlapping nature in NC-OQAM/OFDM, data block length in time domain is much larger, which is clearly shown in Fig. 1. To calculate PAPR, the NC-OQAM/OFDM signal $s(t)$ is segmented into $(M+L)$ intervals. The respective length is T.

It is possible to assume the random nature PAPR in any system that uses the proposed technique. Hence, the complementary cumulative distribution function (CCDF) is used as performance metric. Using this metric, it is also possible to assume that the normal value of the parameter exceeds the threshold.

a. **Transform Formulation**

The Walsh–Hadamard is used in the work for the necessary transformation. The corresponding basis function can be given as

$$\begin{aligned} WH_{ij} &= \frac{1}{\sqrt{M}}(-1)^u \\ u &= \sum_x i_x \bullet j_x \end{aligned} \tag{1}$$

Here, *i and j* are the indices and the order is denoted by M. Also, the symbol (•) denotes a binary AND operation. For demonstration, the respective *WHT* matrix for $M = 2$ is

$$WH_2 = \begin{pmatrix} WH_{1,1} & WH_{1,2} \\ WH_{2,1} & WH_{2,2} \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \tag{2}$$

Alternatively, the matrix can be constructed recursively by defining Hadamard identity $WH_1 = 1$ and WH_M as

$$WHT_M = \frac{1}{\sqrt{2}}\begin{pmatrix} WH_{M/2} & WH_{M/2} \\ WH_{M/2} & -WH_{M/2} \end{pmatrix} \tag{4}$$

Also, it is possible to frame the relations like

$$\begin{aligned} WHT_M &= WHT_M^* = WHT_M^T \\ WHT_M WHT_M &= I_M \end{aligned}$$

where * stands for complex conjugate, T for transpose operation and I_M is an identity matrix of order M. A *WH* transform for a sequence $x(m)$ can be defined as

$$X(k) = \sum_{m=0}^{M-1} WH_{k,m} x_m, \quad k = 0, 1, \ldots, M - 1 \tag{5}$$

where $WH_{k,m}$ is an element of the *WHT* matrix. Figure 2 shows the signal flow graph of this transform. This transform administers the input symbol energy over all subcarriers uniformly because all the elements in the *WHT* matrix have equal magnitude.

b. **ZC Sequence and ZCMT**

The ZC refers to polyphase sequences. The optimum correlation property can be expressed as

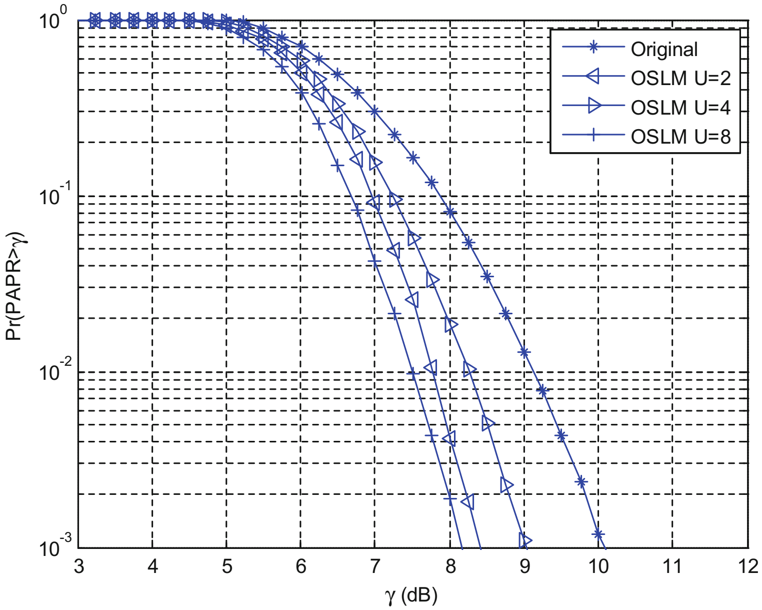


Fig. 2 CCDF curves for OSLM for NC-OFDM/OQAM

$$a_n = \begin{cases} \mathrm{e}^{\frac{j2\pi r}{L}\left(\frac{k^2}{2}+qk\right)} & \text{for} \quad L \quad \text{Even} \\ \mathrm{e}^{\frac{j2\pi r}{L}\left(\frac{k(k+1)}{2}+qk\right)} & \text{for} \quad L \quad \text{Odd} \end{cases} \tag{6}$$

Here, $k = 0, 1, \ldots, L - 1$, $j = \sqrt{-1}$. Also, the q is an integer while the r is any integer which should be the prime to L.

3 Overlapping Selective Mapping for OFDM/OQAM

The direct implementation of the SLM method cannot work with OQAM. To overcome this problem, the SLM method is potentiated the existing techniques in [10, 11] which is called as overlapped selective mapping (OSLM) for case of the OFDM/OQAM. The OSLM method takes into the consideration the overlap between the successive coefficients. The PAPR measurement is carried out over U-dependent coded symbols, and the code that produces the minimum PAPR is selected and stored. This process is repeated for the next symbol, but in this case, the second symbol is the chosen one from the previous selection process.

To generalise this algorithm, it can be described in a step-by-step mode as the following:

1. The code generation: the code vectors $d^{(u)}(u = 1, \ldots, U)$ need to be generated with length equals the number of subcarriers N. And the original symbol $x_{m,n}$ is coded by these code vectors to generate $x_{m,n}^{(i)}$.
2. Creation of the first $2K$ coefficients vectors: the first $2K$ data vectors can be exempted of the multiplication with any code, though they are stored in a matrix of size Nx$2K$:

$$X_{2K} = \left(x_{\mathrm{m},n}^{(1)}\right)_{0 \le m \le M-1, 1 \le n \le 2K} \tag{7}$$

3. Addition of the coding vector: U versions of the coefficients' vector are generated. These U versions are used to create U extended matrices that contain the original matrix X_{2K} and the U editions of the data vector:

$$X_{2K+1}^{(u)} = \left[X_{2K} \; x_{\mathrm{m},2K+1}^{(u)} \right] \tag{8}$$

4. PAPR selection: All the U extended matrices are modulated with OSLM, and then the PAPR is calculated for each output over a window of length equal to the pulse shape length L. This window is located over the symbol of interest, which is the one before the coded vector symbol to contain all the alteration caused by the coded vector symbol addition to the output. The branch with a code index u_{2K+1} that produces the lowest PAPR is selected and stored. Then, the matrix X_{2K} is replaced by $X_{2K+1} = X_{2K+1}^{(u_{2K+1})}$ and maintained a size of N x$2K$.
5. Repeating steps 3 and 4 for the following data: In this step, the coded vector from step 3 is assigned to be the symbol of interest and a new vector is going to be the coded one. So the matrix of consideration in the interval $k > 2K$ is

$$X_{K+1}^{(u)} = \left[X_K \; x_{i,K+1}^{(u)} \right] \tag{9}$$

6. Transmitting the resulting matrix: the resulting data matrix that has the lowest PAPR is then transmitted with the OFDM/OQAM modulator.

4 PAPR Reduction Scheme and Simulation Results

In the proposed SLM, the transformation techniques are employed and then partitioned as subblocks. These transforms are responsible for the decorrelation. They result in better reduction of the PAPR. The block diagram in Fig. 1 is used to demonstrate the proposed scheme. As every subblock is transformed after IFFT, complexity of this scheme increases to $O(UN\log N)$ where $N\log N$ is the complexity of transform and U represents the number of subblocks.

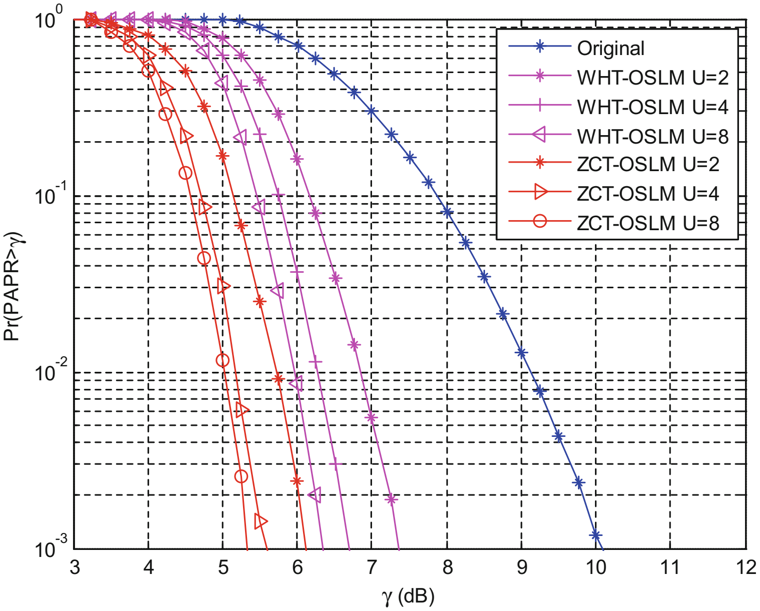


Fig. 3 CCDF curves for the proposed scheme for NC-OFDM/OQAM

The performance of the NC-OFDM/OQAM system is analysed only with OSLM techniques and is shown in Fig. 2. We see that PAPR reduces as the number of blocks (U) increases. CCDF curves have been plotted for $U = 2, 4, 8$. The PAPR of the proposed method is analysed with increased U value and for $U = 2, 4, 8$ the CCDF curves are plotted. Figure 3 shows the simulation results of the proposed method for varying U values, and for comparison the original PAPR of NC-OFDM/OQAM is also plotted with PAPR of 10.1 dB. The two functions, i.e., WHT and ZCT, are merged with OSLM method which are shown to lower this value. It is seen that OSLM combined with ZCT reduces the PAPR to about 7.4 dB. SLM-based WHT gives less performance than ZCT in case of the proposed scheme. The results in Fig. 3 show a better reduction if compared in cases of all transforms.

For the better comparison, the tabulated data in Table 1 can be used. The performance analysis can be inferred from it.

Table 1 Comparison of both the schemes in case of OFDM/OQAM and NC-OFDM/OQAM

OFDM/OQAM						NC-OFDM/OQAM					
Original	WHT	ZCT	OSLM	WHT-OSLM	ZCT-OSLM	Original	WHT	ZCT	OSLM	WHT-OSLM	ZCT-OSLM
10.2	9.1	8.5	8.9	6.3	5.3	10.1	9.2	8.6	8.9	6.4	5.3

5 Conclusion

In this paper, the characteristics of proposed methods in cognitive radio technology are studied. The proposed method appears to be effective in reducing the PAPR of NC-OFDM/OQAM system. The proposed scheme combines the features of both OSLM and all the transforms and provides better reduction in the PAPR of the signal than they can provide when used alone. The combination of both the transformation techniques is possible to witness a case where the symbol observes reduction in PAPR.

References

1. Farhang-Boroujeny, B., Kempter, R.: Multicarrier communication techniques for spectrum sensing and communication in cognitive radios. IEEE Commun. Mag. **46**(4), 80–85 (2008)
2. Amini, P., Kempter, R., Farhang-Boroujeny, B.: A comparison of alternative filter bank multicarrier methods for cognitive radio systems. In: Proceedings of the SDR Technical Conference and Product Exposition (2006)
3. Farhang-Boroujeny, B.: Filter bank spectrum sensing for cognitive radios. IEEE Trans. Sig. Process. **56**(5), 1801–1811 (2008)
4. Jiang, T., et al.: Energy-efficient NC-OFDM/OQAM-based cognitive radio networks. IEEE Commun. Mag. **52**(7), 54–60 (2014)
5. Ye, C., et al.: PAPR reduction of OQAM-OFDM signals using segmental PTS scheme with low complexity. IEEE Trans. Broadcast. **60**(1), 141–147 (2014)
6. Jiang, T., et al.: A novel multi-block tone reservation scheme for PAPR reduction in OQAM-OFDM systems. IEEE Trans. Broadcast. **61**(4), 717–722 (2015)
7. Ghassemi, A., Gulliver, T.A.: PAPR reduction in OFDM based cognitive radio with block wise-subcarrier activation. In: 2012 IEEE International Conference on Communications (ICC). IEEE (2012)
8. Sundeepkumar, V., Anuradha, S.: Adaptive clipping-based active constellation extension for PAPR reduction of OFDM/OQAM signals. Circ. Syst. Sig. Process. 1–13 (2016)
9. Tabassum, S., Hussain, S., Ghafoor, A.: A novel adaptive mode PAPR reduction scheme for NC-OFDM based cognitive radios. In: 2013 IEEE 77th Vehicular Technology Conference (VTC Spring). IEEE (2013)
10. Skrzypczak, A., Javaudin, J.-P., Siohan, P.: Reduction of the peak-to-average power ratio for the OFDM/OQAM modulation. In: IEEE 63rd Vehicular Technology Conference, 2006. VTC 2006-Spring, vol. 4. IEEE (2006)
11. Zhou, Y., et al.: Peak-to-average power ratio reduction for OFDM/OQAM signals via alternative-signal method. IEEE Trans. Veh. Technol. **63**(1), 494–499 (2014)

On the Notch Band Characteristics of CPW-Fed Elliptical Slot

B. Surendra Babu, K. Nagaraju and G. Mahesh

Abstract Slot antennas express multiband characteristics due to its inherent multiresonant structures. Because of its obvious advantages, several shapes of slots are proposed. In this paper, design and simulation of elliptical slot antenna fed with coplanar waveguide (CPW) feed technique for notch band characteristics is presented. The proposed structure is yet another notch band antenna with multiband features which belongs to the slot antenna configuration. The simulation and tuning of the antenna are carried out in computer simulation tool (CST) software which is an excellent.

1 Introduction

The UWB antenna has taken the interest of antenna engineer since the release of the spectrum as unlicensed for civil and commercial applications. It is always a cumbersome task to synthesize antennas for an application without causing interference to other existing application [1–3]. This is possible only if the corresponding has notch characteristics. Hence, in this work, such an antenna which expresses a wide notch band over some portion of the WLAN spectrum is presented. The simulated antenna is analyzed for its UWB features using several radiation parameters. The UWB characteristics are evident from these parameters, and the corresponding notch band

B. Surendra Babu (✉) · K. Nagaraju · G. Mahesh
Department of ECE, Bapatla Engineering College, Bapatla, A.P, India
e-mail: surendrabachina@gmail.com

K. Nagaraju
e-mail: knraju53@gmail.com

G. Mahesh
e-mail: gmahesh033@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_26

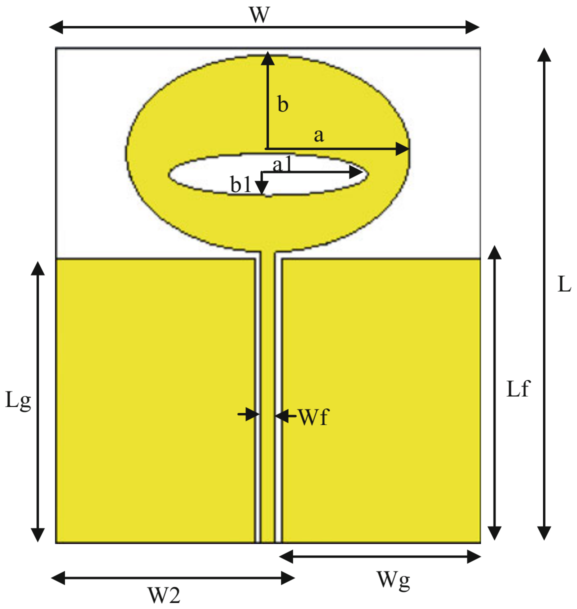


Fig. 1 Geometry of the proposed antenna

features can be inferred [4, 5]. Further, the paper is organized into four sections. Description of the proposed antenna is given in Sect. 2, and the simulation-based experimentation and analyses of the results are given in Sect. 3. Overall conclusion is given in Sect. 4.

2 Proposed Geometry

The typical geometry of the proposed antenna is as shown in Fig. 1. The geometry consists of an elliptical patch on a rectangular substrate. The elliptical patch has dimensions "a" and "b", while the rectangular substrate dimensions are given as "L" and "*W*"; the slot which is also in the shape of ellipse is etched around the center of the patch.

The dimensions of the slot are given as "a1" and "b1". The elliptical patch is excited using complex CPW feed system in order to facilitate the CPW feed system of a thin strip running from the edge of the patch to the substrate end toward the width dimension. The typical feed line is arranged between rectangular ground planes on either sides of the line. The dimension of each rectangular ground plane is "Lg" long and wide by "wg". The length of the strip line which is used as a feed line has a length of "Lf" which is slightly greater than "Lg". The substrate is the FR4 material, and choice is obvious due to its robustness and cost-effectiveness being cheap while the

Table 1 Dimensions of the antenna

S. No.	Parameter	Value (mm)
1	L	35
2	W	30
3	Lg	20
4	Lf	20.5
5	Wf	1.8
6	Wg	14.1
7	$W2$	15.9
8	a	5
9	b	3.5
10	$a1$	3.5
11	$b1$	0.75

width of the patch being 1.6 mm, and the corresponding material dielectric constant is 4.4.

The physical dimensions of the fixed design are listed in Table 1. All the dimensions are in “mm”; the area of the substrate is 10.5 cm^2, while the elliptical patch has largest radius as 5 cm.

3 Simulation Results

The proposed antenna geometry is simulated in CST Microwave Studio using the potential CAD tool integrated into it [4, 5]. The material used for the substrate characterisation is available in the material library and is directly imported from it. The patch is accomplished with PEC boundary conditions, while the entire geometry is enclosed in an air box filled with air dielectrics. The enclosing box surface is assigned with radiation boundary conditions. The radiation characteristics and other radiation parameters are measured along this boundary. After the simulation of the antenna, the min and max frequencies are set for efficient adaptive mesh that is chosen. The final geometry according to the boundary condition is solved.

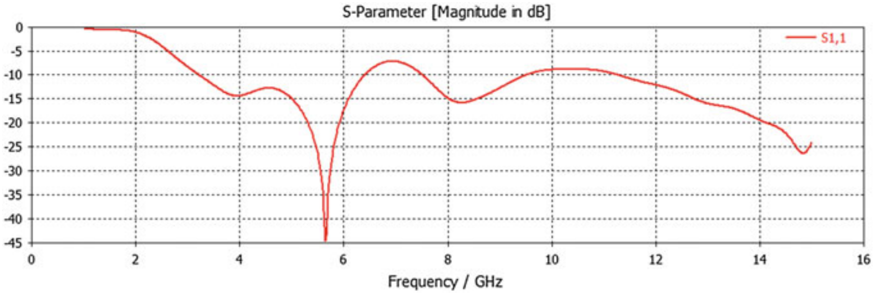


Fig. 2 Reflection coefficient plot

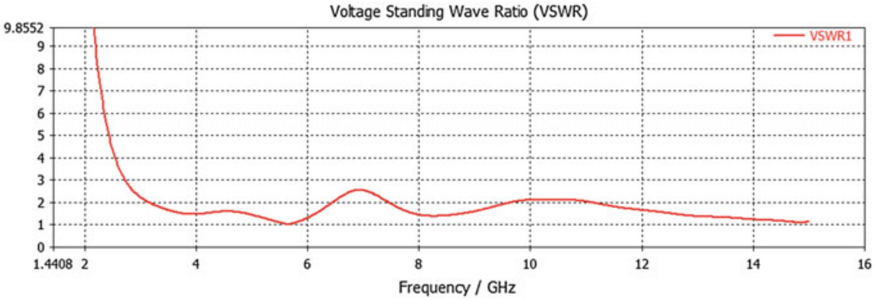
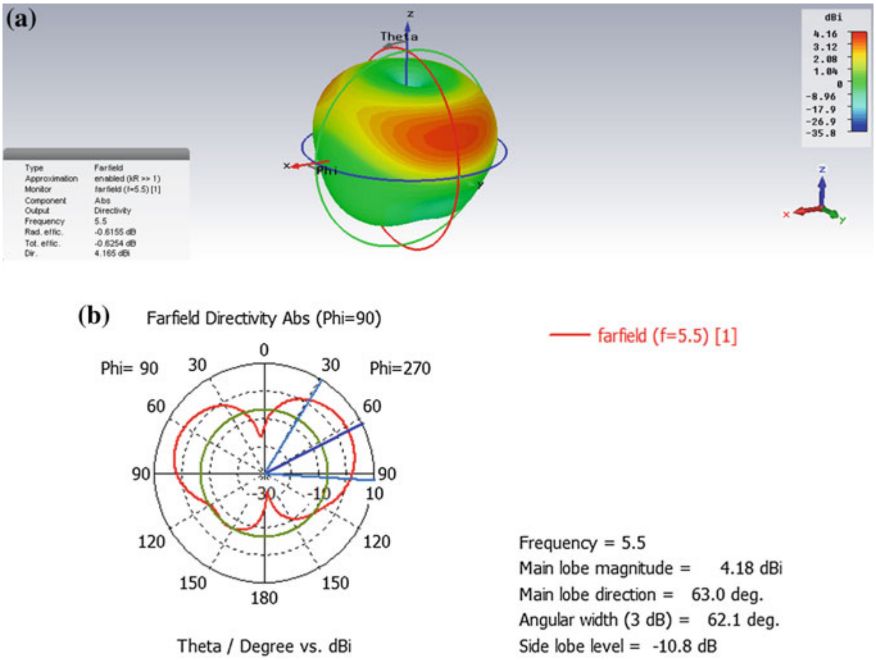


Fig. 3 VSWR plot

Once the geometry is solved, the corresponding filed intensity values are computed from these computed field components, and several radiation parameters can be evaluated. In this work, the analysis of the proposed antenna is carried out for reflection coefficient, VSWR, radiation patterns in 3D and 2D, and current and field distributions plots. On the line, these plots are drawn for the simulated and solved geometry.

The resonant frequencies can be inferred from the S11 plot as shown in Fig. 2. Two bands are ranging from 3.5–6.2 GHz to 7.5–9.5 GHz within the UWB region. In these bands, the corresponding S11 magnitude is well below −10 dB. The same is even evident from the VSWR plot shown in Fig. 3. The magnitude of VSWR is maintained below “2” at these frequencies.



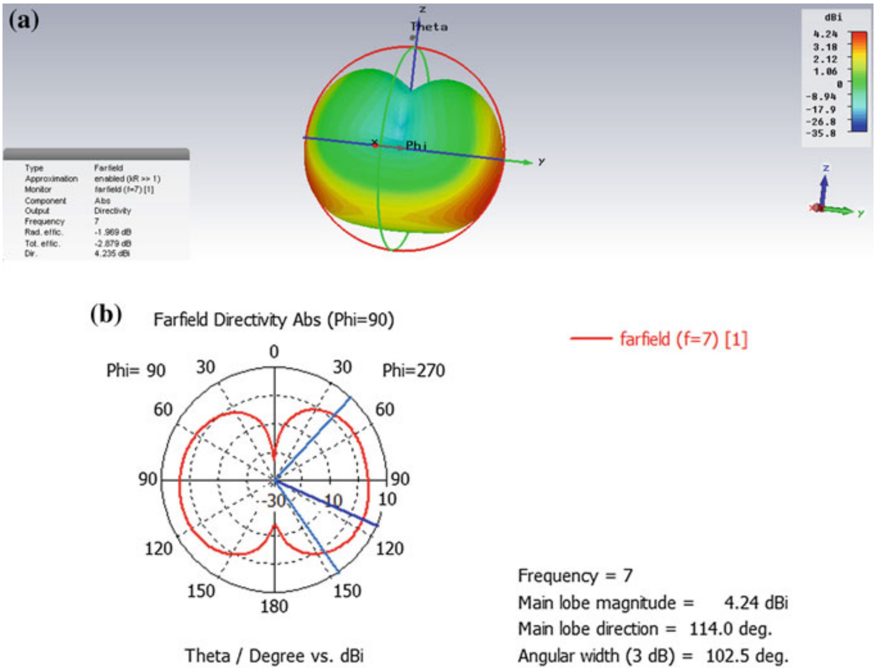
Type	Farfield
Approximation	enabled (kR >> 1)
Monitor	farfield (f=5.5) [1]
Component	Abs
Output	Directivity
Frequency	5.5
Rad. effic.	-0.6155 dB
Tot. effic.	-0.6254 dB
Dir.	4.165 dBi

```
Frequency = 5.5
Main lobe magnitude =     4.18 dBi
Main lobe direction =   63.0 deg.
Angular width (3 dB) =   62.1 deg.
Side lobe level =  -10.8 dB
```

Fig. 4 Radiation pattern plot at 5.5 GHZ in **a** 3D and **b** 2D

For further analysis, the corresponding radiation patterns in 3D and 2D are obtained at 5.5, 7, and 12 GHz as shown in Figs. 4, 5, and 6, respectively. The patterns are similar and express the standard radiation characteristics that of simple patch antenna; it can be inferred from the plot that the lower hemisphere radiation is less than the upper hemisphere.

The E-field and H-field distributions are shown in Figs. 7 and 8. Similarly, the surface current distribution is also obtained as shown in Fig. 9. All the distribution plots are for the 5.5 GHz in the first band and it is evident from the plots that the current density and the corresponding fields are in its high magnitudes in the direction of maximum radiation.



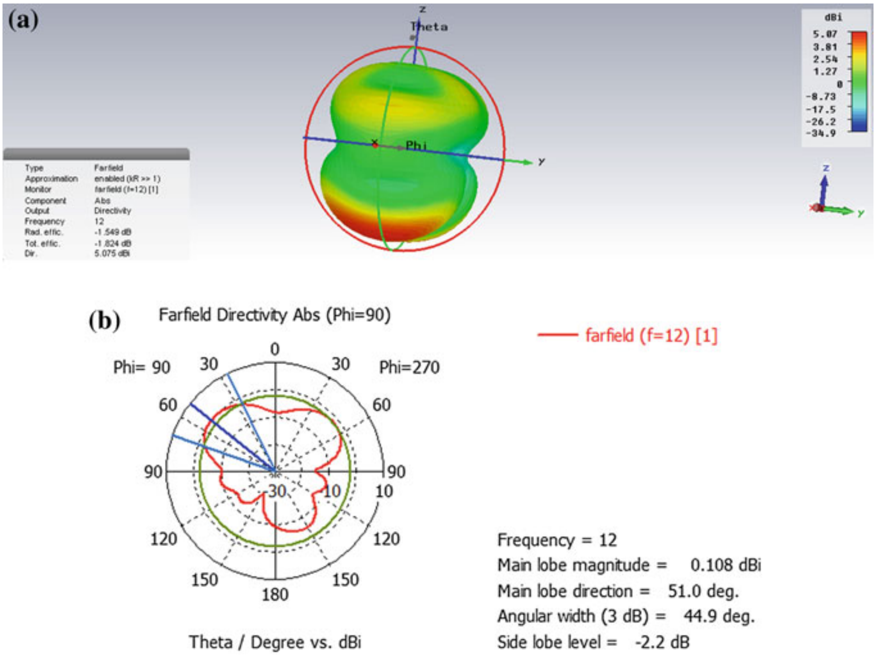
```
Type            Farfield
Approximation   enabled (kR >> 1)
Monitor         farfield (f=7) [1]
Component       Abs
Output          Directivity
Frequency       7
Rad. effic.     -1.969 dB
Tot. effic.     -2.879 dB
Dir.            4.235 dBi
```

```
Frequency = 7
Main lobe magnitude =      4.24 dBi
Main lobe direction =  114.0 deg.
Angular width (3 dB) =  102.5 deg.
```

Fig. 5 Radiation pattern at 7 GHz in **a** 3D and **b** 2D

4 Conclusion

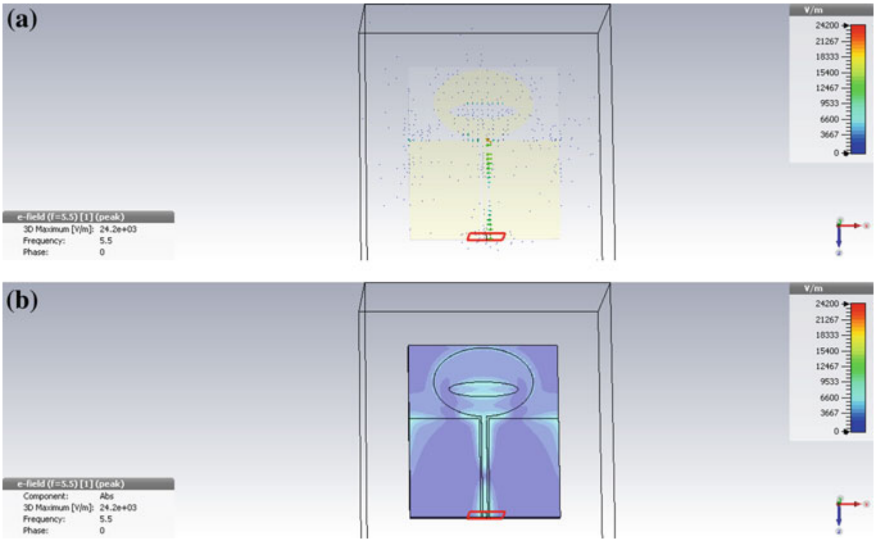
A UWB slot antenna which is CPW-fed is designed and simulated on FR4 substrate successfully. The simulated geometry expressed dual-band characteristics with wide notch bands around the two bands. It also exhibited a wide band of resonance at frequencies outside the UWB range and adjacent to it. The radiation efficiency is at maximum value in the first band. The designed antenna is suitable for UWB applications rejecting some part of the WLAN frequencies.



```
Type            Farfield
Approximation   enabled (kR >> 1)
Monitor         farfield (f=12) [1]
Component       Abs
Output          Directivity
Frequency       12
Rad. effic.     -1.549 dB
Tot. effic.     -1.824 dB
Dir.            5.075 dBi
```

```
Frequency = 12
Main lobe magnitude =    0.108 dBi
Main lobe direction =   51.0 deg.
Angular width (3 dB) =   44.9 deg.
Side lobe level =   -2.2 dB
```

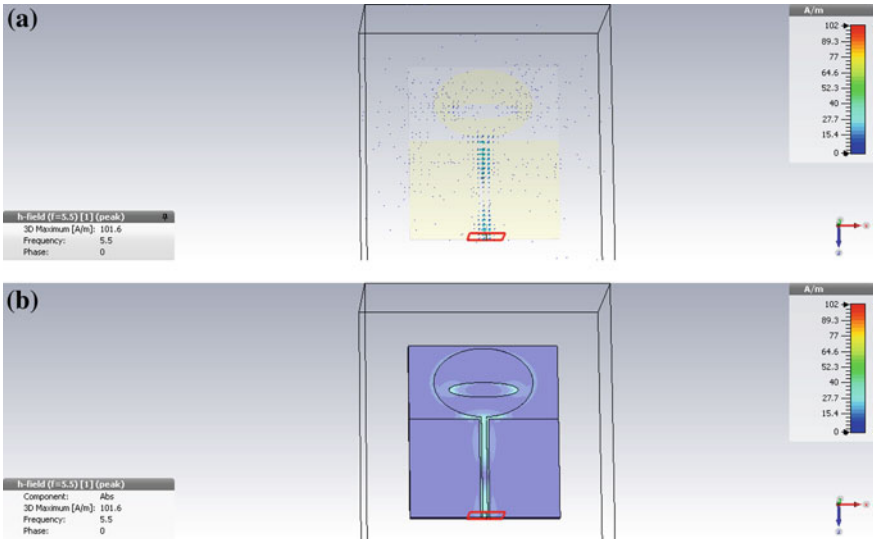
Fig. 6 Radiation pattern plot at 12 GHz in **a** 3D and **b** 2D



```
e-field (f=5.5) [1] (peak)
3D Maximum [V/m]:  24.2e+03
Frequency:         5.5
Phase:             0
```

```
e-field (f=5.5) [1] (peak)
Component:         Abs
3D Maximum [V/m]:  24.2e+03
Frequency:         5.5
Phase:             0
```

Fig. 7 E-field distribution plot



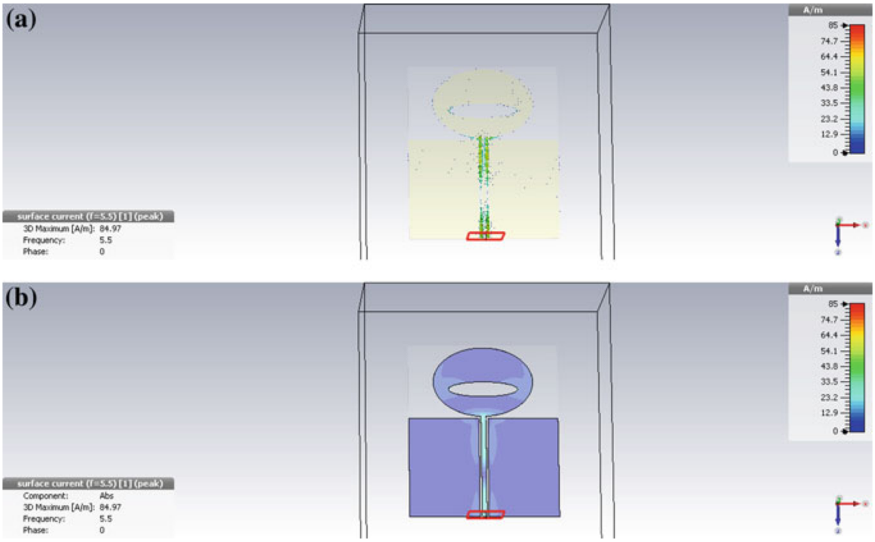
(a)

h-field (f=5.5) [1] (peak)
3D Maximum [A/m]: 101.6
Frequency: 5.5
Phase: 0

(b)

h-field (f=5.5) [1] (peak)
Component: Abs
3D Maximum [A/m]: 101.6
Frequency: 5.5
Phase: 0

Fig. 8 H-field distribution plot



(a)

surface current (f=5.5) [1] (peak)
3D Maximum [A/m]: 84.97
Frequency: 5.5
Phase: 0

(b)

surface current (f=5.5) [1] (peak)
Component: Abs
3D Maximum [A/m]: 84.97
Frequency: 5.5
Phase: 0

Fig. 9 Surface current distribution plots **a** direction and **b** distribution

References

1. Elliott, R.S., Kurtz, L.A.: The design of small slot arrays. IEEE Trans. Antennas Propag. **26**(1), 214–219 (1978)
2. Lee, H.L., Lee, H.J., Yook, J.G., Park, H.K.: Broadband planar antenna having round corner rectangular wide slot. In: Proceedings of IEEE Antennas and Propagation Society International Symposium, 16–21 Jun 2002, vol. 2, pp. 460–463
3. Chen, H.-D.: Broadband CPW-fed square slot antennas with a widened tuning stub. IEEE Trans. Antennas Propag. **51**(8), 1982–1986 (2003)
4. Chakravarthy, V.V.S.S.S., Chowdary, P.S.R., Panda, G., et al.: On the linear antenna array synthesis techniques for sum and difference patterns using flower pollination algorithm. Arab. J. Sci. Eng. (2017). https://doi.org/10.1007/s13369-017-2750-5
5. Deshmukh, A., Desai, A.A., Shaikh, S.A., Lele, K., Phatak, N.V., Ray, K.P.: Elliptical slot cut ultra-wideband antenna. In: 2015 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, pp. 1–6 (2015) https://doi.org/10.1109/conecct.2015.7383870

HDFS Pipeline Reformation to Minimize the Data Loss

B. Purnachandra Rao and N. Nagamalleswara Rao

Abstract The Hadoop is a popular framework. It has been designed to deal with very large sets of data. Hadoop file sizes are usually very large, ranging from gigabytes to terabytes, and large Hadoop clusters store millions of these files. HDFS will use the pipeline process to write the data into blocks. NameNode will send the available blocks list so that pipeline will be created based on the DataNodes having the empty blocks. We can customize the DataNode replacement policy in case of any DataNode failure in the pipeline process using configuration parameters. In this process, write process will be resumed even though there are less number of DataNodes, i.e., even having single DataNode. In single DataNode case, we will lose the data since we have only one copy of data. This paper addresses the issue while having single DataNode in the write operation and taking the pause in write operation until it gets the DataNodes in the pipeline process, and having pause is worthwhile than losing the valuable data if the DataNode fails while write operation is in progress.

1 Introduction

The Apache Hadoop [1] is explicitly designed to handle large amounts of data, which can easily run into many petabytes and even exabytes. Hadoop data files employ a write-once-read-many access model. Data consistency issues that may arise in an updatable database are not an issue with Hadoop file systems, because only a single writer can write to a file at any time. HDFS [2, 3] architecture is implemented in such a way that only one client can write at a time, whereas many clients can read

B. Purnachandra Rao (✉)
Department of Computer Science & Engineering, ANU College of Engineering & Technology, Guntur, India
e-mail: pcr.bobbepalli@gmail.com

N. Nagamalleswara Rao
Department of Information Technology, R.V.R & J.C. College of Engineering & Technology, Guntur, India
e-mail: nnmr3654@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_27

at the same time to reach the data consistency. Hadoop is having the master node called NameNode and it is having namespace. Data will be stored in DataNodes where these are connected to NameNode and periodically sends status report to NameNode. When client applications need to write data to HDFS, they perform an initial write to a local file on the client machine, in a temporary file. When the client finishes the write and closes it, or when the temporary file's size crosses a block boundary, Hadoop will create a file and assign data blocks to the file. The temporary file's contents are then written to the new HDFS file, block by block. After the first block is written, two other replicas (based on the default replication factor three) are written to two other DataNodes in the cluster, one after the other. The write operation will succeed only if Hadoop successfully places all data block replicas in all the target nodes. While writing the data to nodes, there is a possibility of DataNode failures. We can have DataNodes replacement policy using the config parameters. In this, we are addressing the issue of writing data to less number of nodes and the possibility of losing the data.

2 Literature Review

NameNode is having the metadata of the HDFS, and the DataNodes are having application data. DataNode is having blocks. The data will be stored inside blocks. When there is a deletion operation, the hard link from the block will be disconnected. So the data block will remain in the same directory [4]. Blocks on the DataNode contain the file data, and the replication factor depends on the configuration parameter used in the HDFS configuration. Namespace in the NameNode is having information related to blocks and DataNode info of the file. To process the application data very quickly, we can store the frequently accessed data in the cache memory so that the data will be at nearby location, and hence the data access will be faster. Accessing the data without cache will take longer time (milliseconds) compared to accessing the data with cache. We have already observed the performance improvement using cache memory [5]. The large data will be processed by Hadoop. Users will create a key—value pair functionality called MapReduce programming model [6]. Whenever client wants to write data to DataNodes, then client will send a write request to NameNode. NameNode will verify the metadata and find out the empty blocks information, and send the report to output stream. HDFS will create a pipeline based on the replication factor mentioned at the property file. Then, the data will be written to the first block of the DataNode. The data will be copied to another block in the pipeline till the end of the pipeline [7]. Metadata will be stored at NameNode. There is a chance of managing data issues when there is a failure in this NameNode. If NameNode goes down, then the whole system will go down. Metadata replication is one of the solutions to maintain the high availability of the metadata [8]. The DataNodes will send heartbeat to NameNode so that NameNode can understand the availability of data blocks. If the heartbeat is not available within the given time period, then the NameNode will decide that DataNode is dead and that will be replaced by another

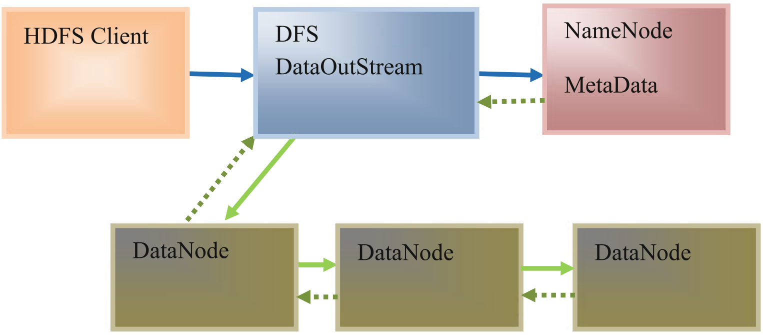


Fig. 1 HDFS write operation replication factor 3

DataNode based on the availability of the DataNodes in the pool [9]. Hadoop stores all log files by moving them from the local file system to HDFS and retaining them there for the duration of the interval configured. Usually, it stores all logs on the nodes where job's tasks have run. We can configure the log aggregation to ensure that you can retain the logs by storing them in HDFS. Log aggregation means that once a job completes, Hadoop will automatically aggregate the job logs from all the nodes where tasks for a job have run and move them to HDFS [10]. Process the log files data to capture the behavior of the system. Data relates to user transactions, user info, system info, and the transaction type. Sometimes, business sales may go down. Customers are simply visiting sites, but not buying the products, customers are abandoning the shopping carts and a sudden rise in support call volume. We can capture all these info in log files. Parsing the log files and getting the behavior of mentioned activities [11]. Figure 1 shows the HDFS write operation. The data packets will be added to the front of the queue if the DataNodes fails so that we can keep them at the safer side. Once the DataNode has been added to pipeline, then it will get new identity after deleting the old DataNodes from the pipeline. The data will be transferred to working DataNodes, while the failed DataNode will be removed [9]. The NameNode will try to arrange the new DataNode since it is under replicated. Even if multiple failures are, there HDFS client will try to continue even if there is only one DataNode. Figure 2 shows the HDFS write operation using replication factor 2, and Fig. 3 shows the HDFS write operation using replication factor 1. Broken red lines in Figs. 2 and 3 mean that there is no communication because DataNodes were down. Table 1 shows the data availability based on the replication factor. If the replication factor is 3, then the data availability is 100%, if it is 2 then availability is 66.66%, and 33.33% if the replication factor is 1. Data availability is 0 if the replication factor is 0, which means it is causing data loss. Please observe the same results in Graph 1. This is the problem in the existing architecture.

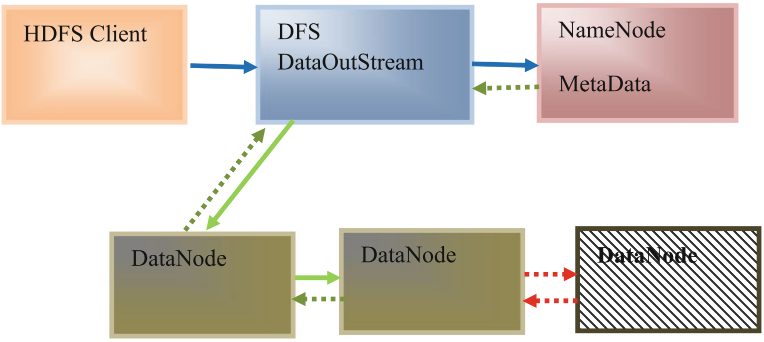


Fig. 2 HDFS write operation replication factor 2

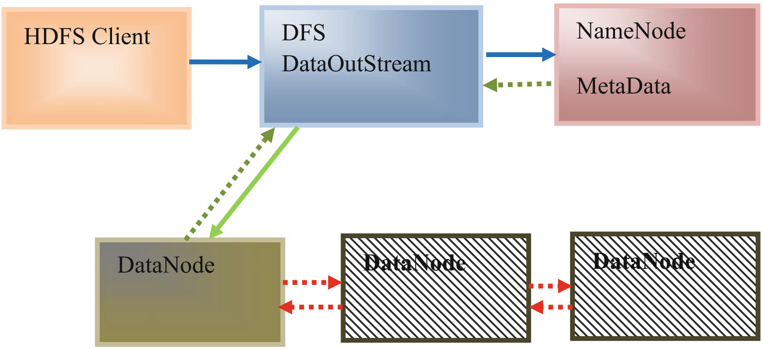


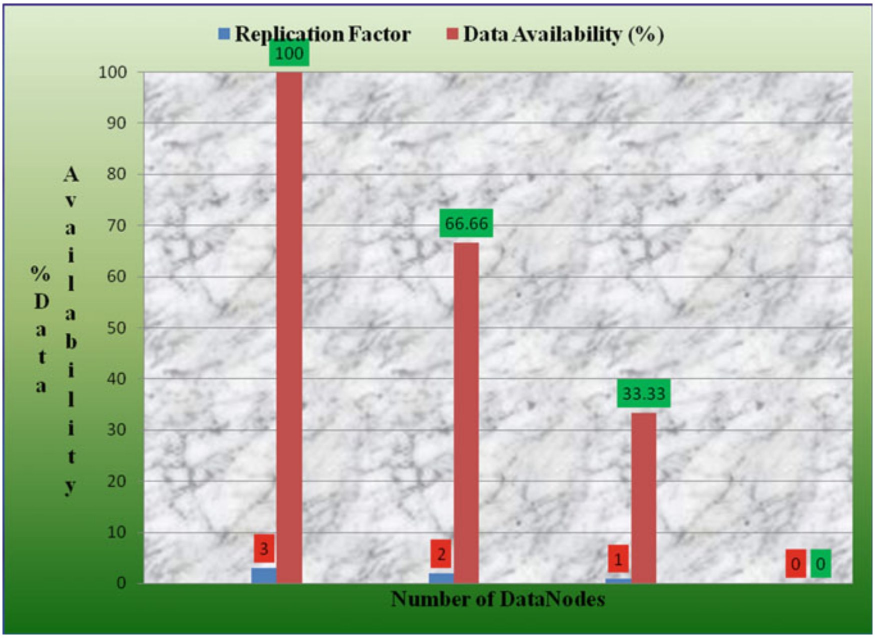
Fig. 3 HDFS write operation replication factor 1

Table 1 Replication factor versus data availability

Replication factor	Data availability (%)
3	100
2	66.66
1	33.33
0	0

3 Problem Statement

While writing data to DataNodes, HDFS will create a pipeline using the info from NameNode. The empty block list will be provided from the NameNode so that it will create the pipeline. Once the write operation is in progress if there is an issue with DataNode, we can configure the DataNode using new DataNode from the available list of nodes. DataNode failure parameter and best effort parameters are available in the Hadoop configuration. Failure parameter will take care about replacing the failed DataNode with the new DataNode, whereas the best effort parameter will try to replace them with the pool DataNode in case of there are very less number of



Graph 1 Replication factor versus data availability

DataNodes at the pipeline. While using the best effort parameter, there are chances of getting exception because of the unavailability of the DataNodes at the pool and the probability of data loss is high. Since we have less number of DataNodes, there is a chance of data loss. Using the property dfs.client.block.write.replace-DataNode-on-failure.best-effort, a client will be able to continue to write even if there is only one DataNode. In this case, we will be having only one replica and if there is an issue for this replica, then there is data loss. This is the problem in the existing environment.

4 Proposal

If there is an issue in the DataNode, then we need to change the data node by configuring the parameters in such a way that it will get replaced by pool DataNodes. The available parameters will replace the existing DataNodes, and the effort parameters will get the DataNode from the pool by force. If there is less availability of DataNodes at the pool, then there is a chance of getting exception at the write operation. And if there are very less number of DataNodes at the pool data loss, chances will be very high, setting the effort parameter to a pipeline with a smaller number of DataNodes. Best effort means client will try to replace the failed DataNode in the pipeline. If the replacement fails also, write operation will continue. In case of single node in the

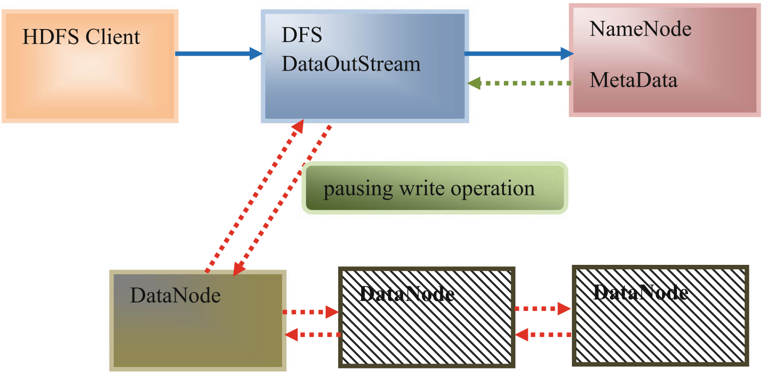


Fig. 4 HDFS write operation suspension

pipeline also, it will start writing the data to the node. This causes the data loss in case of single-node failure. So in case of number of DataNodes are very less, it is better to suspend the data write operation instead of continuing data copying until we get the free block; otherwise, it leads to data loss in case of DataNode failure. By checking the replica number, make sure that the replica is always greater than 1 before proceeding with write operation.

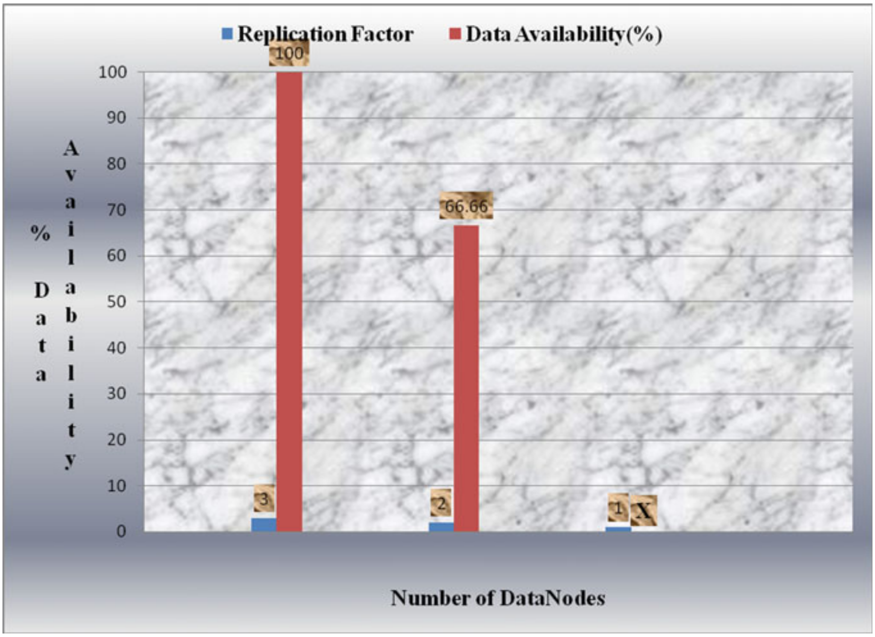
5 Implementation

In the Pipeline recovery process by replacing failed DataNodes, if the number of DataNodes is very less just in case of replacement fails, then better to suspend the write operation for some time and resume the operation once we have enough number of DataNodes in the pipeline process. Figure 4 shows the HDFS write operation process while having very less number of nodes. In the HDFS client write process, the pipeline happened with three number of DataNodes but there was a DataNode failure issue as shown in the diagram, and two DataNodes were down. As per the existing architecture and the available parameters, the write operation can be processed even the pipeline has less number of nodes, say with one DataNode, that is, by enabling failure property parameter and best effort property parameter in the configuration file. But there is a chance of data loss if the one and only one available DataNode goes down. So in this case instead of working with single DataNode in the write operation, it is better we can suspend the operation until we get the number of DataNodes in the pipeline. That is why in the diagram red lines are there from client to first DataNode, i.e., we are suspending the write operation even though it is up and running.

Table 2 shows the simulation results when we apply the pause in the write operation. For replication factor 3, the data availability is 100%, whereas for 2, it is 66.66% and for 1 it is showing as X. When the replication factor is 1, here we are not contin-

Table 2 Replication factor versus data availability

Replication factor	Data availability (%)
3	100
2	66.66
1	X



Graph 2 Replication factor versus data availability

uing the write operation. We need to apply sleep operation (need to decide this factor based on the requirement) in HDFS write operation. So I have mentioned X as the data availability, which means we are not going with write operation if there is only one DataNode. Graph 2 shows the same results, which means it needs to request the NameNode again for getting new DataNodes into pipeline. This is how we can avoid the data loss. But taking the pause from the write operation for couple of milliseconds until it gets the required resources in DataNode pipeline may cause slow down in the performance. It is worthwhile having performance issue than loosing valuable data.

6 Conclusion

In HDFS, write operation will continue even if there are less number of DataNodes. In case any node fails in the pipeline and if the replacement operation fails, then the write operation will be continued which leads to data loss. If a number of DataNodes are very less, i.e., exactly one in that case instead of continuing the write operation, it would be better if we can suspend the write operation for couple of milliseconds (this needs to be determined based on requirement) until we get the additional DataNodes in pipeline process. If we continue with the single DataNode, then there is a chance of data loss if there is an issue with DataNode. In this paper, we have addressed this issue by pausing the write operation for some time (this needs to be determined based on the requirement). The issue with this approach is that performance may go down little bit but it is worthwhile securing the data instead of losing it by facing the little bit performance issue. The future work includes managing the performance as well while securing the data.

References

1. Apache Hadoop. Available at Hadoop Apache
2. Apache Hadoop Distributed File System. Available at Hadoop Distributed File System Apache
3. Scalability of Hadoop Distributed File System
4. Porter, G.: Decoupling storage and computation in Hadoop with SuperDataNodes. ACM SIGOPS Oper. Syst. Rev. **44** (2010)
5. Kakade, A., Raut, S.: Hadoop distributed file system with cache technology. Ind. Sci. **1**(6) (2014). ISSN: 2347-5420
6. Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters. In: Proceeding of the 6th Conference on Symposium on operating Systems Design and Implementation (OSDI’04), Berkeley, CA, USA, pp. 137–150 (2004)
7. Shafer, J., Rixner, S., Cox, A.L.: The Hadoop distributed filesystem: balancing portability and performance. In: Proceedings of IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS 2010), White Plains, NY (2010)
8. Wang, F., et al.: Hadoop High Availability through Metadata Replication. IBM China Research Laboratory, ACM (2009)
9. Tankel, D.: Scalability of Hadoop Distributed File System. Yahoo Developer Work (2010)
10. Alapati, S.R.: Expert Hadoop Administration, Managing, Tuning and Securing
11. Ankam, V.: Big Data Analytics. Published by Packt Publishing Ltd. ISBN 978-1-78588-469-6 (2016)

Short Note on the Application of Compressive Sensing in Image Restoration

Chiluka Ramesh, D. Venkat Rao and K. S. N. Murthy

Abstract Image restoration is a process of reducing the effect of noise and damaged portions in the digital images, and restores images with respective values of neighboring pixels which enhances the image and restores it to original image. To perform this operation filtration, transformation, in-painting, and many other approaches were followed; compressive sensing-based approaches produce best results. In this paper, compressive sensing-based image restoration was studied with different techniques and their comparisons were laid in results section.

1 Introduction

Restoration of an image is the operation of considering a degraded image and evaluating the perfect unique image. Restoration of high-quality images from degraded observations is of growing importance for either esthetic purposes or high-level vision tasks. Estimating the noises present and blur in an image is done to improve the image quality; this is nothing but image restoration. In the course of time with varying environmental and atmospheric conditions, the image gets degraded, so it is necessary to restore the original image using varying image processing algorithms. At present, the technique has widespread application of restoring an image. Image restoration started in 1950s. There are many application domains of image restoration like scientific exploration, legal investigations, filmmaking and archival, image and video decoding, and consumer photography. The main area of application is image reconstruction in radar imaging, radio astronomy, and tomography [1, 2].

Ch. Ramesh (✉) · K. S. N. Murthy
Department of ECE, KL University, Vaddeswaram, Krishna District, Vijayawada, AP, India
e-mail: ch_ramesh_123@yahoo.co.in

K. S. N. Murthy
e-mail: ksnmurty@kluniversity.in

D. Venkat Rao
NIT, Guntur District, Narasaraopet, AP, India
e-mail: dvenky221101@rediffmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_28

The organization of the paper is carried by Sects. 2–4 as mitigation of azimuth ambiguities in SAR images [3, 4], compressive sensing image restoration based on data-driven tight frame, pipelined reconstruction using patch reconstruction, and image restoration using self-similarity.

2 Image Restoration in Synthetic Aperture Imaging (SAR)

In [3, 5], novel system is proposed for relieving azimuth ambiguities in spaceborne strip delineate opening radar (SAR) photos. The azimuth ambiguities in SAR photographs be confined through utilization of an area mean SAR photograph, SAR machine parameters, and a characterized metric got from azimuth reception apparatus design. The characterized metric separates goals lying at spots of ambiguities. The instrument for recuperation of vagueness areas is picked on the premise of a size of equivocalness locales. Compressive imaging approach is enlisted to repair confined vagueness locales (littler districts of interconnected pixels), while grouped zones (phenomenally greater zones of interconnected pixels) are full by method for utilizing model-based aggregate in-painting. The reproduction results on a genuine Terra SAR-X dataset tried that the projected plan can successfully put off azimuth ambiguities and embellish SAR photograph amazingly.

The ambiguities are particularly partitioned into two classes, i.e., azimuth ambiguities and assortment ambiguities. By and large, go ambiguities are more probably appeared as particularly twisted photos, owing to jumbled Doppler rate. Be that as it may, azimuth ambiguities are high-recurrence otherworldly commitments, down changed by methods for examining having practically identical Doppler accuse of regard to the ground destinations lit up by the radio wire essential flap. Consequently, azimuth ambiguities result in additional conspicuous relics in SAR pictures, particularly in waterfront zones, and may be tended to in this paper.

A limited testing of azimuth Doppler alarms presents azimuth ambiguities in spaceborne SAR pictures. Associated alarms are created when collapsing of Doppler frequencies toward the central piece of radio wire test in Doppler recurrence area. The vague alarms could be dislodged symmetrically in azimuth toward the best possible and left of the genuine objective part. Azimuth ambiguities make decreased sign clamor proportion (SNR) and result in problematic visual incredible of the SAR picture. In extraordinary, an amazingly extreme power objective may likewise bring about obviously solid phony destinations in low-profundity homogeneous legacy areas (e.g., ocean bottom) at areas dislodged by method for both range shifts and azimuth. The image restoration based on image ambiguities flowchart is shown in Fig. 1.

Restriction and relief of ambiguities are the techniques connected to reestablish a picture created by radar. The duplicated enthusiasm of studies organized in direct opposite issues may cause exceptionally green recuperating calculations in not so distant future. We are sure that, abusing cutting edge inquire about in those zones, the proposed structure is viable and mathematically productive choice for moderation

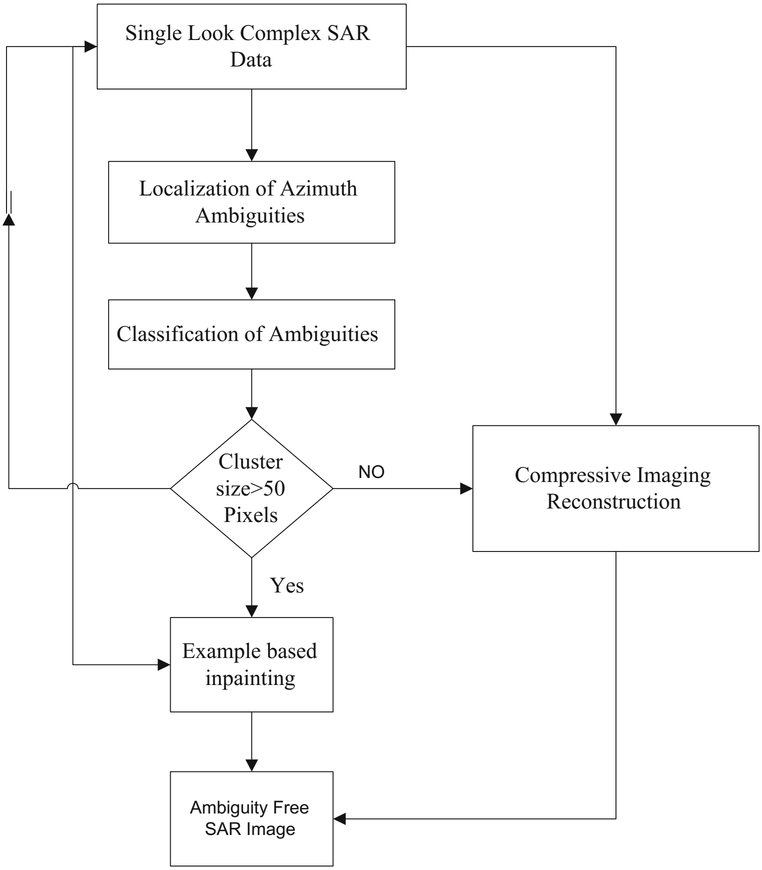

Single Look Complex SAR Data
Localization of Azimuth Ambiguities
Classification of Ambiguities
Cluster size>50 Pixels
NO
Yes
Compressive Imaging Reconstruction
Example based inpainting
Ambiguity Free SAR Image

Fig. 1 Image restoration based on image ambiguities

of azimuth ambiguities in spaceborne strip outline, yet best PSNR was not obtained as same ambiguities different pictures.

3 Compressed Sensing Restoration of Image-Based Data-Driven Tight Frame

However, now compressive detecting (CS) rebuilding with tight frame is connected to the pictures. In [6], it demonstrates that repetitive flag representation, e.g., rigid body, plays out an essential capacity in packed detecting picture recuperation. Keeping in mind the end goal to get an astounding scanty outline, one has tried persisting endeavors to seek after tight casings.

In spite of the fact that there are some tight edges underneath which a sort of photographs has an amazing inadequate estimate, some other type of depictions might not have meager guess because of the pictures' superb qualification fit as a fiddle. This [7] paper offers a particular compacted detecting picture recovery strategy in light of the measurement-driven multiscale rigid casing. This technique infers a discrete multiscale tight casing machine versatile to the one of a kind photo from entering packed detecting picture. Such a versatile tight body creation conspire is actualized to packed detecting photo recuperation.

It can be seen that the motivation clamor identifier assumes a basic part in the methodologies of those changed middle sort channels. On the off chance that one noisy pixel is appeared in light of the fact that the commotion free pixel by means of the finder, it'll be uncorrected inside accompanying strategies; in any case, on the off chance that one clamor detached pixel shows up in light of the fact that the loud pixel, which can be filled in by utilizing its surrounding pixels [7]. In this correspondence, a solitary strategy meant for dispensing with salt-and-pepper commotion is projected. The considerable pick up of the system is the disentanglement of the clamor location. The method recognizes loud pixel by methods for judging regardless of whether its dark esteem breaks even with the most or least incentive inside dynamic range. Albeit some clamor-free pixels will likewise be thought about as loud pixels, it will affect the general execution of the approach inconsequentially.

4 Pipelined Image Reconstruction Using Patched Reconstruction

Looking for the scanty estimate of a given picture assumes an essential part in compacted detecting picture reclamation undertakings. The settled wavelet tight edges have been generally used to reestablish the compacted detecting picture, using the pictures' sparsity under the frames [8].

Piece-based absolutely arbitrary photograph testing is combined with a projection driven a compacted detecting recovery that supports sparsity in the territory of directional changes all the while with a clean recreated picture. Both contourlets, and in addition complex-esteemed dual-tree wavelets, are mulled over for his or her very directional portrayal, while bivariate shrinkage is custom fitted to their multiscale deterioration structure to offer the considered essential sparsity requirement [8].

Smoothing is performed through a Wiener sift through joined into iterative anticipated Landweber compacted detecting recovery, yielding quick reproduction. The proposed procedure yields previews with a fine that fits or surpasses that delivered by a prevalent, however, computationally rich, method which limits general variety. Moreover, recreation quality is fundamentally best in class to that from various remarkable interests based on absolute calculations that do exclude any smoothing [8, 9]. The pipelined-based reconstruction using de-blocking is explained in Fig. 2.

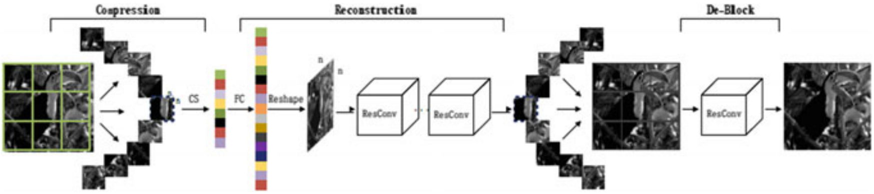


Fig. 2 Pipelined reconstruction using patched reconstruction and de-blocking [9]

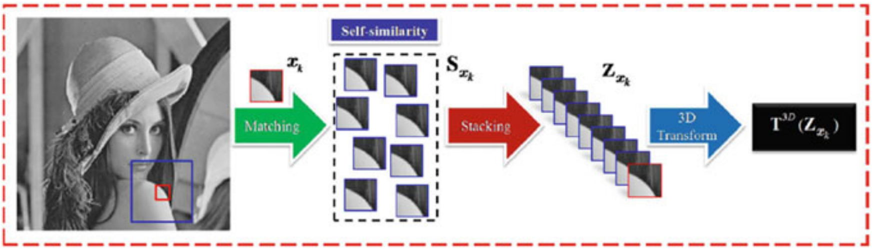


Fig. 3 Image restoration using self-similarity

5 Image Restoration Using Self-Similarity

A progression of strategies had been proposed to remake a photo from compressively detected irregular size, be that as it may; the majority of them have unreasonable time unpredictability and are unseemly for fix-based completely compacted detecting catch, as a result of their genuine blocky ancient rarities inside the recuperation results. In this paper, we show a non-iterative picture remaking come nearer from fix-based compressively detected arbitrary estimation. Our strategy abilities fell systems construct absolutely with respect to lingering convolution neural group to take in the stop-to-stop full picture recuperation, which is fit for remaking picture fixes and disposing the blocky effect [9] (Fig. 3).

Picture restoration is the operation of taking a worsen photograph and surveying the right, interesting photograph. At first, the assortment is isolated from a stuck scene and composed with the expression reference and is stacked aggregately. At residual, the pics are restored using SDL estimation. The PSNR attributes are taken a gander at to be higher than an ordinary state of all weight methodologies. The purpose of word reference learning is to discover a territory in which some preparation data surrenders a lacking depiction. In this technique, the examples are taken under the Nyquist charge. Regardless, in one of kind occurrences, a dictionary that is prepared to fit the data can basically embellish the sparsity, which has bundles in insights breaking down.

CS is a recently rising methodology and a strikingly considered inconvenience in sign and picture preparing, which proposes a spic-and-span structure for synchronous inspecting and pressure of inadequate or compressible markers at a rate essentially underneath the Nyquist cost [10]. Perhaps, planning a viable regularization day and age mirroring the photograph meager earlier data plays a basic position in CS photo reclamation.

As of late, adjacent smoothness and nonlocal self-similitude have achieved prevalent sparsity before CS photograph rebuilding. A versatile curvelet thresholding model is produced in this paper, looking to adaptively get rid of the bothers showed up in recuperated photos at some phase in CS recuperation methodology, forcing sparsity. Moreover, another sparsity degree known as joint versatile sparsity regularization (JASR) is set up; this implements each adjacent sparsity and nonlocal 3-d sparsity in rebuild area, simultaneously. At that point, a solitary strategy for high-consistency CS picture rebuilding through JASR is proposed—CS-JASR. To effectively clear up the proposed relating improvement inconvenience, we lease the split Bergman cycles. Broad trial results are specified to confirm the sufficiency and viability of the proposed approach contrasted and the cutting edge ultra-current systems in CS picture reclamation [10, 11].

6 Conclusion

The intrinsic houses of nonlocal self-similarity and local smoothness of natural snapshots are considered from the angle of facts at the identical time. Experimental results on three packages: photograph in-painting, picture deblurring, and combined salt-and-pepper and Gaussian noise elimination have proven to facilitate the proposed algorithm to achieve huge overall performance developments over the modern day schemes and reveal first-rate convergence property. Future work consists of the research of the facts for natural photos at more than one scale and orientations and the extensions on ramification of programs, including deblurring of an image with combined Gaussian and impulse noise and video recovery obligations.

References

1. Zhang, J., Zhao, D., Xiong, R., Ma, S., Gao, W.: Image restoration using joint statistical modeling in a space-transform domain. IEEE Trans. Circ. Syst. Video Technol. **24**(6), 915–928 (2014)
2. Yang, J., Sha, W.E.I., Chao, H., et al.: Multimed. Tools Appl. **75**, 6189 (2016). https://doi.org/10.1007/s11042-015-2566-9
3. Chen, J., Iqbal, M., Yang, W., Wang, P.B., Sun, B.: Mitigation of azimuth ambiguities in spaceborne stripmap SAR images using selective restoration. IEEE Trans. Geosci. Remote Sens. **52**(7), 4038–4045 (2014)
4. Avolio, C., Mario, C., Di Martino, G., Antonio, I., Flavia, M„ Giuseppe, R., Daniele, R., Massimo, Z.: A method for the reduction of ship detection false alarms due to SAR azimuth ambiguity. In: 2014 IEEE Geoscience and Remote Sensing Symposium (2014)
5. Xie, Z., et al.: Restoration of sparse aperture images using spatial modulation diversity technology based on a binocular telescope testbed. IEEE Photon. J. **9**(3), 1–11 (2017)

6. Huang, S., Zhu, J.: Removal of salt-and-pepper noise based on compressed sensing. Electron. Lett. **46**(17), 1198–1199 (2010)
7. Chunhong, C., Gao, X.: Compressed sensing image restoration based on data-driven multi-scale tight frame. J. Comput. Appl. Math. **309**, pp. 622–629 (2017)
8. Mun, S., Fowler, J.E.: Block compressed sensing of images using directional transforms. In: 2010 Data Compression Conference, Snowbird, UT, pp. 547–547 (2010)
9. Nie, G., Fu, Y., Zheng, Y., Huang, H.: Image restoration from patch-based compressed sensing measurement arXiv:1706.00597 (2017)
10. Zhang, J., Zhao, D., Zhao, C., Xiong, R., Ma, S., Gao, W.: Image compressive sensing recovery via collaborative sparsity. IEEE J. Emerg. Sel. Top. Circ. Syst. **2**(3), 380–391 (2012)
11. Eslahi, N., Aghagolzadeh, A.: Compressive sensing image restoration using adaptive curvelet thresholding and nonlocal sparse regularization. IEEE Trans. Image Process. **25**(7), 3126–3140 (2016)

An Automated Big Data Processing Engine

Leonid Datta, Abhishek Mukherjee, Chetan Kumar and P. Swarnalatha

Abstract In such continuously changing era when large chunks of data are generated at every moment, data analysis is performed for business predictions. The processing of such data is very difficult to be handled in serialized manner. To avoid such constraint, we opt for parallel processing. The term big data refers to the large and complex data chunks which cannot be processed using day-to-day processing software because of their limitations. And also, in the existing environment where the big data are processed, the system is controlled by an admin, i.e., the processor is not automated. In this system, we propose to develop an automated engine that will receive the dataset and the requirements for the output as input from the user, and the engine will process that chunk of data without involvement of an admin according to the need of the user and the output will be generated.

1 Introduction

In this paper, we propose to develop an engine which solves the problem of analyzing large chunks of data in an efficient manner. This system will be efficient and effective enough to make the process of data storage and data processing hassle free and will be abstract and robust enough for users to keep and analyze the large amounts of private data. Parallel processing technique will be used for a more optimized handling of these chunks of data. In parallel processing technique, the large number of data processes to be performed on the whole dataset are divided into a number of small

L. Datta (✉) · A. Mukherjee · C. Kumar · P. Swarnalatha
VIT University, Vellore, India
e-mail: leoniddatta@gmail.com

A. Mukherjee
e-mail: scobbyabhi9@gmail.com

C. Kumar
e-mail: chetan_tulsyan@yahoo.com

P. Swarnalatha
e-mail: pswarnalatha@vit.ac.in

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_29

parts and processed in parallel. Finally, all the outputs of the processes are merged into a single output which can be treated as the final output of the processing. The proposed system is built on Hadoop. The Hadoop software library is a framework which permits distributed processing of data across large clusters of computers using various platforms for data analysis, data storage, etc. The main benefit of using Hadoop is that it uses commodity servers for data storage and processing which are relatively cheaper than dedicated servers, and hence Hadoop reduces the cost required for server maintenance. The analysis tasks of the proposed system will be based primarily on MapReduce algorithm. It is a related implementation for processing and generating large datasets that is amenable to a broad variety of real-world tasks [1]. MapReduce is used because it takes the input data and maps them into key–value pairs that help in optimizing all future operations. The proposed system will have a file drop system which can be used by the user to upload the required files for a process and then select the type of analysis to be performed as per requirement. This system will consist of an engine having user logins where the users can submit queries and keep a track of the status of the query as required by the user for the processes which have been submitted.

2 Related Work

Much before the generation of chunks of data which are now used for analysis, Ithiel de Sola Pool used the sent words for finding a pattern for the flow of information [2]. It is used to analyze the flow of words. This concept was later modified for business analysis and prediction by finding different statistical parameters and patterns. Due to the continuous growth of data generation, storing cost increased steadily [3]. Along with that, the complexity of data analysis also increased. To reduce this complexity, MapReduce was introduced which supports parallel processing and simplified analysis of large clustered data. In recent times, WEKA has been introduced which provides a general-purpose environment for automatic classification, regression, clustering, and feature selection—common data mining problems in various areas related to machine learning-based research problems [4]. These are used for classification and prediction using the analysis, and it provides access to the database using the database query method. In case of spatial dataset also, MapReduce has been used for analysis [5]. But in all the cases the analysis is done with the help of an admin. The proposed system tries to set a link between the HDFS and an http portal through HttpFS and thus resulting in a system that can analyze the data and make prediction without direct involvement of an admin.

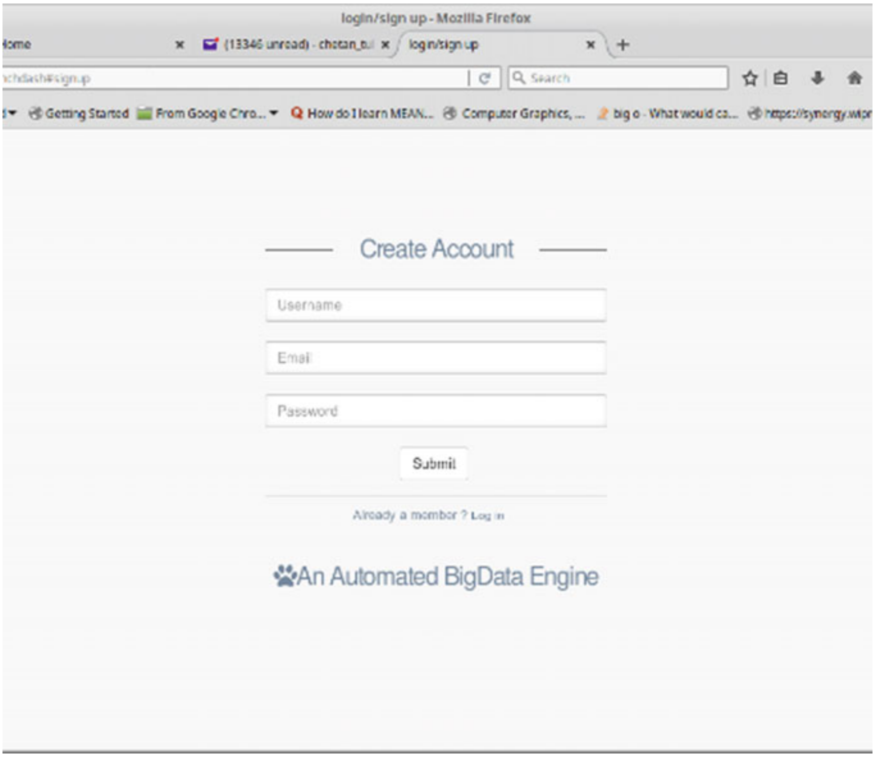

login/sign up - Mozilla Firefox
Create Account
Username
Email
Password
Submit
Already a member ? Log in
An Automated BigData Engine

Fig. 1 Login page for user

3 Methodology

The system consists of two main modules: first being an engine which serves as a portal for the user and the other being data processing which processes the request received from the portal and carries out the requested processing task. The web module consists of various sub-modules like login and signup for users, profile updating (refer Fig. 1) for users, and data query submission. The user can log in using these modules and update the profiles if required. The user can submit queries as per requirement and track progress of the process which is continuously updated by the user. When the user logs into the portal, he gets the option to update profile or submit a query. The user submits the input file and selects the type of operation, e.g., word count or filtering data or clustering, in the portal. The main advantage of this system is that the admin cannot see the input dataset. It can only see the preferred type of operation. Hence, it provides a good security to sensitive data and also maintains the privacy of the dataset given by the user. When the query is submitted, the situation is next handled by the data processing modules.

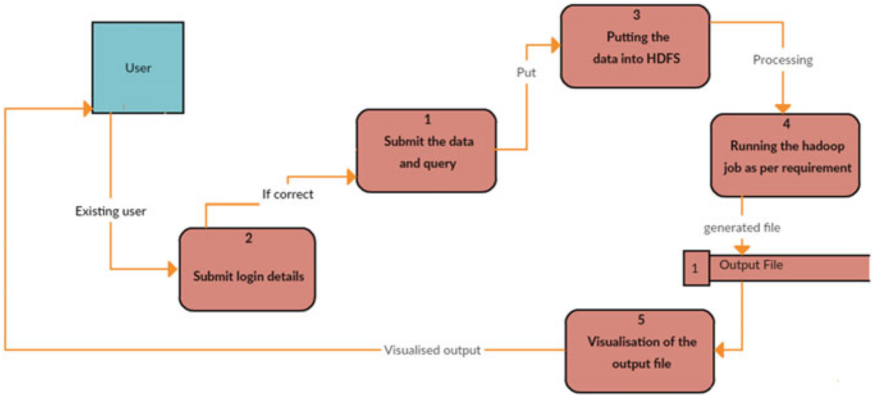

User
Existing user
2
Submit login details
If correct
1
Submit the data and query
Put
3
Putting the data into HDFS
Processing
4
Running the hadoop job as per requirement
generated file
1
Output File
5
Visualisation of the output file
Visualised output

Fig. 2 Data flow diagram

The data processing module has many sub-modules, namely, data cleaning, data aggregation, data filtering, data organization, and finally data visualization. All of these sub-modules are managed in Hadoop Distributed File System (HDFS) which is basically a distributed file system. However, the differences from other distributed file systems are many. HDFS can handle faults and errors in a much better manner as compared to other distributed file systems, and it can be used with commodity low-cost hardware systems. Hence, it can store large amounts of data in a distributed fashion. HDFS runs on a master–slave concept, i.e., there is a master node which holds the metadata of the data contained in the slave nodes of HDFS. The master node is called the NameNode whose main task is to store the locations, size of data, number of copies created, etc. which are required for easy access of the files which are present in the slave nodes, i.e., the DataNodes. This makes the file system more robust and secure since one node is exclusively allotted for the details of the data which have been stored in DataNode and since it stores copies of the data blocks, the efficiency of access of the data gets more optimized in a way that if one DataNode is not accessible or is having high latency in the fetching the data from HDFS other DataNodes which are holding the copies of the data which are to be accessed can be used. Parallel computing works on any computing or processing environment in which many parts or the execution of processes are done simultaneously, i.e., large problems can be divided into smaller problems for processing and then at the end, they are merged for producing output. Parallel processing reduces the processing time drastically because more than one portion of the data is being processed at the same time. Figure 2 can be referred for more clear understanding of the flow of processing that how the input dataset is being processed in a cycle and the prediction or the expected output is given back to the user.

The sampling sub-module performs the task of collecting randomized chunks of data from the input data as is available so as to get a basic insight into the data and to perform a few preprocessing tasks. This has many use cases in the corresponding sub-modules one of which is the data organization sub-module where the range of data is required. This is meant for the ordered sorting use case where data ranges are defined for data partitioning purposes since according to these ranges the input data to be sorted is sent to various subprocesses where each partition is sorted according to the required parameter and after that the sorted results of all these individual partitions are combined to get the sorted data as a whole.

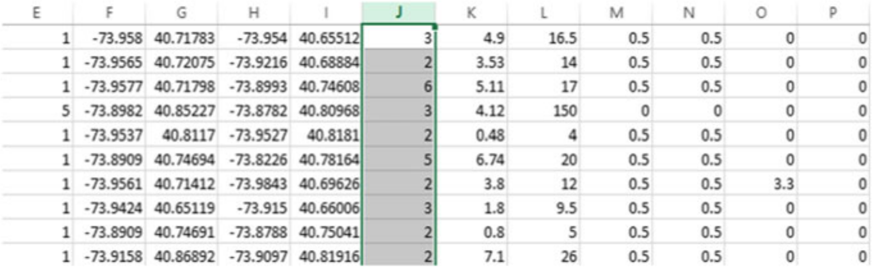
Another use of the sampling sub-module is finding the delimiter of the input data which is required to use structured data for analysis purposes.

The data cleaning sub-module performs the task of checking the attributes of the data and verifying whether the available data in each attribute matched with the domain of that attribute. Sometimes, due to errors at the source of data, null values are available instead of values which data sources, irrelevant to originating from inside or outside of the organization, may need cleansing because errors might create problems while processing data. To work upon such data, numerical null values or string null values may be filled in a specific manner to aid the process. In case of numerical data cleaning, the data cleaning sub-module calculates the average of each column and fills all the null numerical values with the average obtained from the average calculation process.

The data aggregation sub-module gives the user a feature to get brief insights into the data before performing actual analysis on the data to the aid the processing to optimizing the requested process as well as to know about some of the statistical measures on the attributes of the data. It deals with calculation of various statistical parameters like mean, median, standard deviation, variance, range, etc. Mean and median serve as the measure of center, while variance and standard deviation serve as the measure of distribution. For calculating these statistical parameters, MapReduce has been used as basic algorithm and depending on the parameter the algorithm has been implemented. Like for mean calculation, map phase reads the input file line by line and generates the key–value pair. Also, it performs sort and shuffle which leads to the sorting of column values according to keys. Reduce phase has ready list of values. So, it performs summation of the values and the summation is divided by total number of pairs. Thus, mean is calculated.

Another important aspect of data aggregation in case of data having discrete values is that we need to know the frequency of occurrence among the dataset as a whole to which tells us about the distribution of the data in its domain. Data aggregation also deals with finding most occurring words in a document which are later used for tagging documents, and hence used for further analysis as per requirement one which is TF-IDF algorithm that is mainly used for such tasks.

The filtering layer sub-module extracts certain chunks of data on the basis of some applied function or randomly extracting chunks which vary in size and dimension. The data sampling sub-module is basically a type of filtering where random samples are taken from the data. It is done by means of randomly allotting numbers evenly to the various lines in the data and checking for the modulo of the line number matching within a range of numbers. Figure 3 describes an output sample of filtering module. If a larger sample is required, we take a large range which signifies more amount of data and if a smaller range is required we take a smaller range of data



E	F	G	H	I	J	K	L	M	N	O	P
1	-73.958	40.71783	-73.954	40.65512	3	4.9	16.5	0.5	0.5	0	0
1	-73.9565	40.72075	-73.9216	40.68884	2	3.53	14	0.5	0.5	0	0
1	-73.9577	40.71798	-73.8993	40.74608	6	5.11	17	0.5	0.5	0	0
5	-73.8982	40.85227	-73.8782	40.80968	3	4.12	150	0	0	0	0
1	-73.9537	40.8117	-73.9527	40.8181	2	0.48	4	0.5	0.5	0	0
1	-73.8909	40.74694	-73.8226	40.78164	5	6.74	20	0.5	0.5	0	0
1	-73.9561	40.71412	-73.9843	40.69626	2	3.8	12	0.5	0.5	3.3	0
1	-73.9424	40.65119	-73.915	40.66006	3	1.8	9.5	0.5	0.5	0	0
1	-73.8909	40.74691	-73.8788	40.75041	2	0.8	5	0.5	0.5	0	0
1	-73.9158	40.86892	-73.9097	40.81916	2	7.1	26	0.5	0.5	0	0

Fig. 3 Sample output file for data filtering where all values within a range (2–6) are organized in one file

to be matched. One of the most important functionalities in this layer is finding of the top *n* (any positive integer) values based on a metric as specified by the user. Another functionality in this layer is that of finding all distinct values present in a string of parameters so that all repetitions are filtered out. So this layer takes in data and gives sampled based on user-based custom conditions or random conditions. In selection, the command has been designed as “E a b” where a is the column number from where the values will be checked and *b* is the value to be compared with the values of the column, i.e., “E 1 2” will select the tuples where values at column 1 are equal to 2. Similarly, “R a b c” is designed in a way that it will select the values from column number a which are in greater than or equal to b and less than c. The data organization sub-module takes input data and based on the user requirement operates on a single or multiple datasets, which performs requested operations. This is specially done to ease the task of distributive systems and to bifurcate certain chunks of data into different parts based on some classification parameter. Binning performs this task of segregating data. Data organization also takes in multiple data and performs various operations like data joins where multiple datasets are joined on basis of certain parameters as specified by the user. This joining works efficiently in cases where the required column is not in a single table. So, through different types of joins, the required data column is fit in a single table. The left outer join returns all the tuples from the left table and the tuples from right table that match records from the right table. It returns NULL from if match is not found. Similarly, in case of right outer join, it returns all the tuples from the right table that are reserved and only the matching values from the left table are returned. If normal join is performed, then number of lost tuples are huge. So, these left and right outer joins are performed. But, this also leads to loss of data which is not suggested in data analysis. Figure 4 describes the flowchart of filtering sub-module.

So, full outer join is preferred if more than one columns are taken into consideration for analysis. For performing the join, map phase generates the primary key and its corresponding column value as the key–value pair. The reduce phase takes the values of those tuples for which the primary keys match and then joining is performed. Depending on the type of join, the algorithm selects the table. Like, in case of left outer join, in map phase, the left table’s primary key and its corresponding

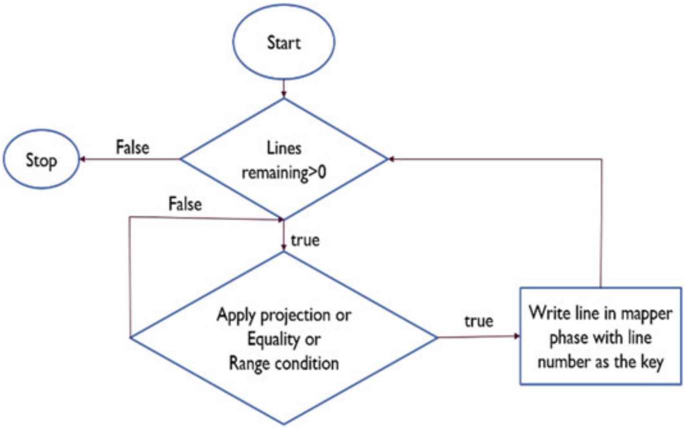
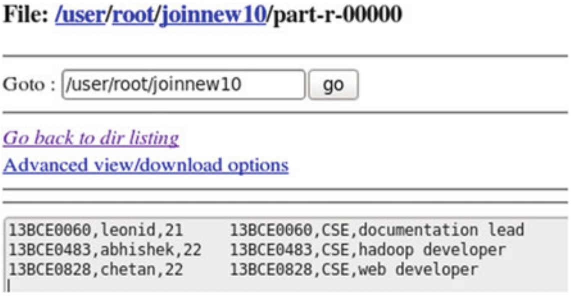

Start
Lines remaining>0
False
Stop
False
true
Apply projection or Equality or Range condition
true
Write line in mapper phase with line number as the key

Fig. 4 Flowchart for data filtering by applying data equality, range, and projection conditions

Fig. 5 Sample of performed join operation



File: /user/root/joinnew10/part-r-00000

Goto : /user/root/joinnew10 go
Go back to dir listing
Advanced view/download options

```
13BCE0060,leonid,21     13BCE0060,CSE,documentation lead
13BCE0483,abhishek,22   13BCE0483,CSE,hadoop developer
13BCE0828,chetan,22     13BCE0828,CSE,web developer
```

column value are taken as key–value pair. Reduce phase takes out the matching pairs and joined output is generated. Figure 5 shows a sample for the joining operation.

The data classification sub-module helps in categorizing data into right category. Classification is an important data mining technique with broad applications to classify the various kinds of data used in nearly every field of our life [6]. This module helps in getting a clear view of the data. It takes the dataset and trains the machine (hardware) using various algorithms to classify the data into particular category. This is generally used to predict class variables of the input testing data from the given training data, and hence helps to categorize the data into classes on basis of the training dataset as specified by the user. This can be done by various techniques one of which is the k-nearest neighbor algorithm where the processing of classification is done to classify the whole dataset into a set of k classes with high intra-class variance and high inter-class similarity. The objective of the visualization layer is that it helps in getting a better understanding of the data. This layer makes the picture more clear to a general user. After running a specified algorithm on the input dataset, an output file will be generated which will be fed into the software named Tableau along with the cluster details and port number. This will generate a graphical representation of our output data for improved understanding. The type of representation can also be customized according to the user demands.

4 Future Work

In the proposed system, future work may include the regression techniques which help in the cases where the data flow follows certain equations. This is mainly for the two-dimensional data where support vector machine can be used in case of *n*-dimensional data. The proposed system through analysis of data can lead to only prediction of certain values. But if we want to include prediction of probabilistic values of an event with probability of happening and probability of not happening, the naïve Bayes classifier can be included for that.

5 Conclusion

The proposed system utilizes the quality of analyzing large chunks of data through HDFS efficiently and features of a web portal that is connected to a big data processing framework. The architecture of the proposed system is designed in a way that it can receive the dataset from the user and the requirements for the analysis of the dataset from the user. It deals efficiently with the dataset, and processing is done at the back end of the web portal which is capable of handling a large number of service requests simultaneously at a single time. Starting from cleaning of the received dataset up to visualization of the analyzed data, they help in analysis through different aggregation methods and the technique of MapReduce to calculate required statistical parameters also. These features can be used with the help of Internet connection only, and thus it makes the system automated and also it matches the industry standard which is required in this era of continuously growing efficiency.

References

1. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Commun. ACM **51**(1), 107–113 (2008)
2. Ithiel de Sola, P.O.O.L.: Technologies of Freedom. Harvard University Press (1983)
3. Morris, R.J., Truskowski, B.J.: The evolution of storage systems. IBM Syst. J. **42**(2), 205–217 (2003)
4. Frank, E., Hall, M., Trigg, L., Holmes, G., Witten, I.H.: Data mining in bioinformatics using Weka. Bioinformatics **20**(15), 2479–2481 (2004)
5. Eldawy, A., Mokbel, M.F.: Spatialhadoop: A mapreduce framework for spatial data. In: 2015 IEEE 31st International Conference on Data Engineering (ICDE), pp. 1352–1363. IEEE (2015)
6. Patil, T.R., Sherekar, S.S.: Performance analysis of Naive Bayes and J48 classification algorithm for data classification. Int. J. Comput. Sci. Appl. **6**(2), 256–261 (2013)

Predictive Analysis of Stocks Using Data Mining

G. Magesh and P. Swarnalatha

Abstract There are 60 major stock exchanges around the world with a total value of $69 trillion. Stocks are traded almost daily. Stock data is available on the Internet right from the beginning. Prediction of stock market is an attractive topic for researchers of different fields. Before the advent of machine learning and data science, stock market movement was primarily analyzed using statistical and technical factors. Now with the help of machine learning techniques, it is possible to accurately identify the stock market movement. Various machine learning techniques like support machine vectors, random forests, gradient boosted trees, etc. have been successfully used in the past to predict stock prices.

1 Introduction

Stock prediction is multibillion-dollar industry. News blogs twitter tweets reflect the mood of the buyer. They are the first one to reflect how market will move the coming day. All the models constructed so far seem to ignore the fact that people are using Internet to communicate and express themselves a lot more. So a comprehensive model to analyze stock markets with the help of machine learning, tweets, news blogs, etc. is required. Various researches have been conducted in this field. There are various companies which have proprietary trading algorithms to automate stock market trading. Most of the algorithms focus on using a statistical model or a combination of statistical model with machine learning techniques. None of them uses the fact that Twitter tweets and news articles combined can be a powerful tool for stock market prediction. When an unpredictable event happens like when a company misses its target, our prediction algorithm may not always be correct. In this case, tweets can come in

G. Magesh (✉)
School of Information Technology and Engineering, VIT University, Vellore, India
e-mail: magesh.g11@gmail.com

P. Swarnalatha
School of Computer Science and Engineering, VIT University, Vellore, India

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_30

handy. Twitter is one of the first platforms where news are shared publicly. We can avoid certain unseen errors in our model using this twitter analytics support.

2 Relevant Works

Random forest, gradient boosted trees, and support machine vectors were used to predict short-term market movement. Data of Morocco Telecom of Casablanca stock market was used. Data was collected in a 10-min interval. Two methods mean decrease accuracy and mean decrease impurity were used in particular [1].

The data was obtained from Bloomberg terminal for 3M Stock of New York Stock Exchange. The time frame chosen was from September 1, 2008 to August 11, 2013 [2]. Totally, 1471 data points were chosen. They used 16 technical features for their prediction model. They divided the model into two parts—Next-day model and long-term model. In next-day model, supervised machine learning techniques were used. Techniques like logistic regression, Gaussian discriminant analysis, quadratic discriminant analysis, and SVM were used.

More than 755 million tweets were analyzed. Lexicon-based approach gave importance to eight basic emotions in tweets. Stocks from Dow Jones Industrial Average and S&P 500 were used. Two major machine learning algorithms were used—Neural networks and support vector machine. Stock price data from Yahoo Finance was obtained [3].

A study was on major stock indices—DAX, DJIA, FTSE-100, HSI, and NASDAQ [4]. They used neural networks with an accuracy of around 60%. There was a model using Weka by Nikola Milosevic to analyze stocks over a 1-year period.

We give confirmation of the helpfulness of misusing on the web content information in stock expectation frameworks. Running trials in which the models are prepared with various mixes of components extricated from the past conduct of stock costs, or mined from the online message sheets. Proof recommends that it is conceivable to extricate prescient data from stock message sheets [5].

We offer a precise examination of the utilization of profound learning systems for securities exchange investigation and expectation. Our investigation endeavors to give a far-reaching and target appraisal of both the favorable circumstances and disadvantages of profound learning calculations for securities exchange examination and expectation [6].

The paper analyzes the open doors in and conceivable outcomes emerging from huge information in retailing, especially along five noteworthy information measurements—information relating to clients, items, time, (geospatial) area, and channel. A significant part of the expansion in information quality and application conceivable outcomes originates from a blend of new information sources, a shrewd utilization of measurable instruments, and area learning consolidated with hypothetical bits of knowledge [7].

As monetary joining and business association's increment, organizations effectively connect with each other in the market in helpful or aggressive connections. To comprehend the market arrange structure with organization connections and to explore the effects of market arrange structure on stock division execution, we propose the development of an organization relative system in view of open media information and part collaboration measurements in view of the organization arrange [8].

Anticipating securities exchange returns is a testing undertaking because of the perplexing idea of the information. This investigation builds up a nonspecific approach to foresee day-by-day stock value developments by sending and incorporating three information diagnostic forecast models: versatile neuro-fluffy surmising frameworks, fake neural systems, and bolster vector machines [9].

Stock cost changes are accepting the expanding consideration of financial specialists, particularly the individuals who have long-haul points. The present investigation plans to evaluate the consistency of costs on Tehran Stock Exchange through the utilization of counterfeit neural system models and essential segment examination strategy and utilizing 20 bookkeeping factors [10].

This paper demonstrates an assessment of the adequacy of the feeling investigation in the stock expectation assignment through a vast scale test. Looking at the exactness normal more than 18 stocks in 1-year exchange, our technique accomplished 2.07% preferred execution over the model utilizing verifiable costs as it were [11].

This paper breaks down the prescient capacity of the particular esteem deterioration entropy for the Shenzhen Component Index in light of various scales. It is discovered that the prescient capacity of the entropy for the file is influenced by the width of moving time windows and the auxiliary break in securities exchange [12].

This paper endeavors to offer a more extensive meaning of enormous information that catches its other extraordinary and characterizing attributes. The quick development and reception of huge information by industry has jumped the talk to famous outlets, constraining the scholarly press to get up to speed. Scholarly diaries in various orders, which will profit from an important talk of huge information, presently cannot seem to cover the point [13].

In this paper, both specialized and crucial examinations are considered. Specialized investigation is finished utilizing recorded information of stock costs by applying machine learning, and key examination is finished utilizing web-based social networking information by applying opinion examination [14].

In this paper, we demonstrate to utilize web search tool information to gauge close term estimations of monetary markers. Illustrations incorporate vehicle deals, joblessness claims, travel goal arranging, and purchaser certainty [15].

In this investigation, we have built up a candle graph examination master framework, or an outline translator, for foreseeing the best securities exchange timing. The master framework has examples and guidelines which can anticipate future stock value developments [16].

3 Method

3.1 Data Source

The stock data was obtained from yahoo finance. Stocks from various top stock indices like NIFTY100, S&P 500, and Russell 2000 were obtained. We obtained stocks from start of 2013 to December of 2016. The main financial indicators obtained were

Market capitalization—Total number of shares/Present share price,
Ask—The price the owner is willing to sell,
Bid—The price the buyer is willing to pay,
Average Day volume—Average number of shares sold per day over a particular period,
Book value—The amount of money a common shareholder would get in case the company would liquidate,
P/E Ratio—Current share price per share earning,
Earnings per share—Company's profit shared per common share,
Price/Book Ratio—Current closing price of stock/price of stock per latest book value,
Price/Sales Ratio—Market Cap/Most recent revenue,
52 Week Low—The lowest this particular stock went in the past year,
52 Week High—The highest this particular stock went in the past year,
Current Ratio—Ratio of current assets to current liability, and
Sales Growth—Sales Growth over 6-month periods.
We scrapped a total of 4500 stock data.

For twitter data analysis, we created an app on Twitter. All the relevant keys were saved and the app was given all the required permission. A Java package Tweet4J [5] was used to obtain tweets from twitter. Tweets were obtained with respect to the company in question. The number of tweets to be retrieved was given. The tweets were stored in JSON format in our database.

3.2 Prediction Technique

Our main aim is to classify data as to whether the stock will rise or not. So basically it turns out to be a classification algorithm. Since we have already scraped data already for the past 4 years, we can easily classify data whose value increase after a specified period of time. We took a time period and checked if after 6 months if the stock increased in value. We wrote a script for this so that this could be automated. We labeled data as good whose value increases after 6 months as "SAFE" and the others "UNSAFE". Equal number if "SAFE" and "UNSAFE" data were used in our experiment. We clean the data and assign values of –MAX for missing values.

We used Weka toolkit [5] for our machine learning model. We used various models like C4.5 decision trees, support vector machine, random trees, random forest, and naïve Bayes. We used the following steps to train the model using Weka.

- The model is trained using the "SAFE" "UNSAFE" attribute we added to our dataset.
- We used only 70% of our data.
- The rest 30% was used to test our data.

First, all the features extracted from Yahoo were used. Then, manual feature selection was used to enhance our accuracy. We removed a feature and checked if it increased accuracy. This process was done continuously to make the accuracy as high as a possible.

4 Results

The results obtained while using all features are given in Table 1.

A few features selected manually are display in Table 2.

As seen above, the algorithm that performed best was random forest with precision, recall, and F-score with 72% accuracy. When the same random forest was performed with manual feature selection, we got an accuracy of 77%. The features that were selected were

Table 1 Table with all features

S. No.	Algorithm	Precision	Recall	F-score
1	C4.5 decision trees	0.64	0.64	0.64
2	SVM	0.62	0.62	0.62
3	Random tree	0.65	0.65	0.65
4	Random forest	0.72	0.72	0.72
5	Naïve Bayes	0.52	0.52	0.52

Table 2 Table with all selected features

S. No.	Algorithm	Precision	Recall	F-score
1	C4.5 decision trees	0.68	0.68	0.68
2	SVM	0.64	0.64	0.64
3	Random tree	0.69	0.69	0.69
4	Random forest	0.77	0.77	0.77
5	Naïve Bayes	0.55	0.55	0.55

- Market cap,
- Book value,
- PE ratio,
- Earnings per share,
- Price/Books ratio, and
- Price/Sales ratio.

The tweets were returned with a sentimental score attached to it. The total sentimental score of the entire tweet sets was calculated. If the tweets had a negative score meaning that the current trend is negative about the company for some reason, we alert the user immediately. This would make the user cross-check our prediction in rare cases.

5 Conclusion

In this paper, we used machine learning algorithms to predict stock market movement. With all the 13 features selected, we got the highest accuracy of 72% using random forest model. By selecting features manually, we improved our accuracy to 77%. There were certain features which were not required for stock analysis. Those features removed gave an improved accuracy to our model. Our model performed with 77% F-score while performing correctly on 77% instances. This model can be used by people trading stocks to get advice for their investment or to verify if their decision is correct. The tweet analysis provides an added layer of protection in case something suddenly happens which our model cannot predict.

References

1. Labiad, B., Berrado, A., Benabbou, L.: Machine learning techniques for short term stock movements classification for moroccan stock exchange. In: 11th International Conference on Intelligent Systems: Theories and Applications (2016)
2. Dai, Y., Zhang, Y.: Machine Learning in Stock Price Trend Forecasting. Stanford University Research (2015)
3. Porshnev, A., Redkin, I., Shevchenko, A.: Machine learning in prediction of stock market indicators based on historical data and data from Twitter sentiment analysis. In: 13th International Conference on Data Mining Workshops (2013)
4. Phua, P.K.H., Zhu, X., Koh, C.H.: Forecasting stock index increments using neural networks with trust region methods. s.l. IEEE, pp. 260–265 (2003)
5. Zarandi, M.H.F.: A type-2 fuzzy rule-based expert system model for stock price analysis. Expert Syst. Appl. **36**(1), 139–154 (2009)
6. Smailović, J., Grčar, M., Lavrač, N., Žnidaršič, M.: Predictive sentiment analysis of tweets: A stock market application. Lecture Notes Computer Science (including Subseries Lecture Notes Artificial Intelligence Lecture Notes Bioinformatics), vol. 7947 LNCS, pp. 77–88 (2013)
7. Sandhiya, V., Revathi, T., Jayashree, A., Ramya, A., Sivasankari, S.: Stock Market Prediction on Bigdata Using Machine Learning Algorithm, vol. 7(4), pp. 10057–10059 (2017)

8. Paranjape-Voditel, P., Deshpande, U.: A stock market portfolio recommender system based on association rule mining. Appl. Soft Comput. J. **13**(2), 1055–1063 (2013)
9. Oztekin, A., Kizilaslan, R., Freund, S., Iseri, A.: A data analytic approach to forecasting daily stock returns in an emerging market. Eur. J. Oper. Res. **253**, 697–710 (2015)
10. Nichante, V., Patil, P.S.: A Review : Analysis of Stock Market by Using Big Data Analytic Technology, pp. 305–306 (2008)
11. Nguyen, T.H., Shirai, K., Velcin, J.: Sentiment analysis on social media for stock movement prediction. Expert Syst. Appl. **42**(24), 9603–9611 (2015)
12. Navale, P.G.S., Dudhwala, N., Jadhav, K., Gabda, P., Vihangam, B.K.: Prediction of Stock Market Using Data Mining and Artificial Intelligence, vol. 6(6), pp. 6539–6544 (2016)
13. Nann, S., Krauss, J., Schoder, D.: Predictive analytics on public data—the case of stock markets. In: Proceedings of 21st European Conference Information Systems, pp. 1–12 (2013)
14. Kim, Y., Jeong, S.R., Ghani, I.: Text opinion mining to analyze news for stock market prediction. Int. J. Adv. Soft Comput. Appl. **6**(1), 1–13 (2014)
15. Kanade, V., Devikar, B., Phadatare, S., Munde, P., Sonone, S.: Stock market prediction: using historical data analysis. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **7**(1), 267–270 (2017)
16. Junqué De Fortuny, E., De Smedt, T., Martens, D., Daelemans, W.: Evaluating and understanding text-based stock price prediction models. Inf. Process. Manag. **50**(2), 426–441 (2014)

Efficient Load Balancing Algorithm for Task Preprocessing in Fog Computing Environment

A. B. Manju and S. Sumathy

Abstract Focus and research on fog computing environment is increasing in recent days. Orchestrating the resources in the fog computing environment and distributing the task with the help of simple load balancing algorithms improve the task processing in fog environment. Fog resources include end users resources, networking resources, and cloud resources as well, in which networking resources take the central control over the fog nodes at particular location. These control nodes attempt to reduce the burden of cloud as well as improve the task processing efficiency by distributing the load across the fog nodes evenly on controlling the fog nodes based on availability of the nodes. The objective of this work is to evenly distribute the load across the available fog nodes and reduce the response time of the task processing. The results have been verified using cloud analyst tool in which the proposed approach is compared with round-robin algorithm in terms of response time. The results show that the response time has improved substantially in the proposed approach.

1 Introduction

Fog computing was basically a term that was introduced by Cisco Systems, to initially define the fog resources which were available at the network edges such as routers, switches, and base stations. Fog computing overcomes the disadvantages of cloud computing such as latency, location awareness, security, and so on [1]. Though cloud computing is available at flexible charges with all services, fog computing outstands cloud with all its benefits such as extreme low latency, improved security, location awareness, and improved quality of services for real-time applications. Without burdening the network bandwidth, the real-time data can get communicated with local datacenters which are available close to the users [2].

A. B. Manju (✉) · S. Sumathy
Vellore Institute of Technology, Vellore 632014, Tamil Nadu, India
e-mail: manju.ab@vit.ac.in

S. Sumathy
e-mail: ssumathy@vit.ac.in

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_31

The real-time application that needs low latency data processing has rapidly increased with evolving technology. Unlike cloud data processing, these applications demand short response time. In addition, location-aware data is produced in IoT applications. Fog computing is evolving with all these advancements, which is otherwise called as local cloud [3]. In spite of local datacenters distributed close to the user's end, the request handling by the datacenters needs to be orchestrated in order to increase the efficiency of the task processing. While there are many open issues in integrating cloud and IoT devices, the load at cloud can be efficiently handled with the help of fog devices.

Load balancing in fog environment improves the performance of the task processing at fog nodes. Fog nodes include end users resources and network resources. Fog nodes are distributed across the geographical locations; they preprocess the task and forward the task to cloud datacenter. Reliable network resources at particular locations are assigned as control nodes. Allocating the users request to appropriate datacenters is the responsibility of the service broker. In this approach, the service broker is the network resource at each location. User's request must be processed according to the priority as well as based on the cost of processing the task [2]. Some applications need faster response irrespective of the cost of processing, whereas some applications need to process the task at minimum cost. When there are huge number of requests arising from the same geographical location, the datacenter near that location get overloaded, and hence there must be some load balancing policies in order to distribute the task without overwhelming at the same datacenter. In case of fog computing, the scenario remains the same. If the user's task at particular location gets overloaded at that particular fog clusters, then proper load distribution rules must be incorporated for even distribution of load. As of fog environment, the real-time processing and faster response is the ultimate target, and the load balancing policy must not take long time. The load balancing policy must be suitable for fluctuating resource change in the fog environment.

2 Literature Review

2.1 Load Balancing in Fog

Load balancing policies applied in cloud can be extended to fog environment, which should be modified according to the characteristics of the fog resources and task. The fog networking and its architecture have been given more importance in research field as the number of IoT devices is increasing [4]. The task offloading, with energy optimization in fog networks, is analyzed in work [5], where the energy is optimized by segmenting the arriving task into processing, offloading, and networking-related task. With fog networking, many augmented reality applications are being efficiently built and being evolving which demands at most real-time communication [5]. The fog nodes have very less processing capabilities. Fog nodes reduce the burden of cloud task processing. Function of networking device is forwarding the data packets according to the predefined rules. Fog node that is formed with the help of user's end resources can be managed with the help of these networking devices at each location. Unlike fog environment where the nodes are not reliable, cloud environment has stable resources that allow formulating standard load balancing policies. Range-wise busy checking 2-way balanced algorithm is proposed [6]. It has three phases in choosing the appropriate cloudlet which may prolong the period of choosing the cloudlets. Load balancing by predicting the load at each edge node and at the neighboring nodes [7]. The work has been carried out in real-time environment by taking the data from the edge nodes of single server. On applying the load balancing at edge nodes, they have achieved up to 93% of accuracy. The results vary when same policy is achieved at different geographical locations; hence, accuracy cannot be same at all places. A load balancing policy based on the user's priority and characteristics of the datacenter is discussed [8]. The work has been segmented into three parts, such as considering users priority, determining the datacenter characteristics, and finally deciding which node to assign the task. These constrains can be best suitable in terms of changing resources at the edge computing environment. Cluster-based server-based fog resource provisioning approach has been proposed [9], in which group of networking devices acts as clusters to balance the load across the network. In the work carried out by author in [10], they have stated that processing the task at fog nodes will reduce the cost of the task processing. They have formulated the method by determining the number of requests to be processed at each fog node. There are certain parameters to be considered for calculating the average latency of the task offloading such as queue length, and a number of waiting task to be served in the queues are discussed [11]. The work proposed in [12] deals with improving the quality of service of users in allocating the task to appropriate VMS. This work has reduced the completion time of the task which is based on improvements done in min-max algorithm.

2.2 Min-Min Algorithm

Min-min algorithm has been applied in cloud and grid computing widely. A modified min-min algorithm with priority to the users is discussed [13]. A hybrid min-min algorithm called genetic-based min-min is proposed in [14], in which the size of instructions in the task is considered. The best fit task for the resources is selected by fitness based on length of the task. In [15], min-min algorithm works in the modified way and the authors have partitioned the task into groups and the group of task which has larger makespan is selected for execution first.

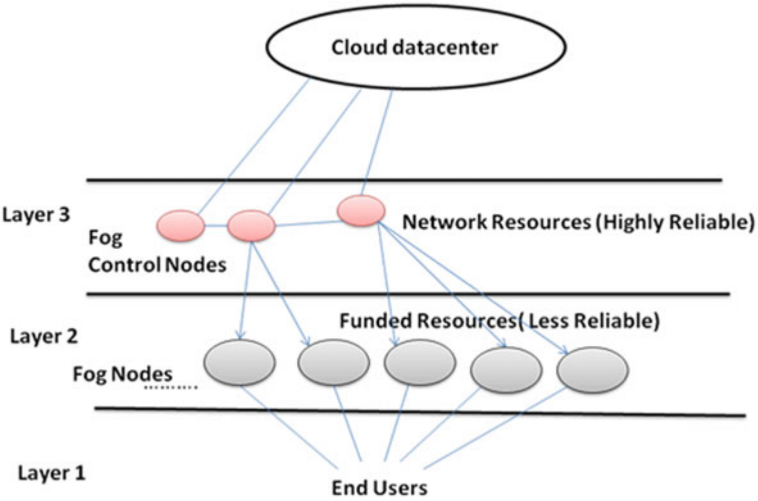


Fig. 1 Fog-based load balancing architecture

3 Proposed Approach

The architecture model explained in this approach is shown in Fig. 1. The four-layered architecture gives the details regarding the fog computing environment existing as a proxy between the end users and the cloud service providers. With the help of fog devices (laptops, server, routers, and so on), the load at the cloud can be reduced. In this approach, the reliable networking resources such as routers, switches, and base stations are assigned as the control nodes. The fog load balancing architecture is segmented into four-layered architecture such as end users, unreliable fog nodes, reliable network resources, and cloud datacenters. End users are those who request for task processing. End users can be any Internet-connected device. Unreliable fog nodes can be laptops, mobile phones, tablets, connected vehicle, and so on. These resources from a particular geographical location form as clusters to act as a fog datacenter at that location. The resources at this cluster dynamically change based on the availability of the resources at that location. Reliable networking resources act as a controller node. It receives updates about the number of available idle and busy nodes and is responsible for forwarding the task to appropriate nodes based on minimum latency.

3.1 Forwarding the User’s Task to Appropriate Node

When user’s task is received at the fog device, fog device preprocesses the task before forwarding to the cloud environment. User’s task is always be forwarded to nearest fog device for preprocessing. The controller node at each location gets the heartbeat updates about the number of idle resources available at that location, as well as the neighboring locations.

At each location, the unreliable fog nodes form a cluster under a controller node which maintains metadata about the number of idle and busy nodes. Each cluster gets task from that particular location. Controller node forwards the task to appropriate fog node based on the availability of the resources and also based on the requirements of the task. Under each cluster, min-min load balancing algorithm is used for balancing the load. If the local nodes are busy, controller will check for the neighbor node availability and forward the task to neighboring nodes. If neighboring nodes are busy, then the task is directly forwarded to the cloud for processing.

3.2 Constrain Based Min-Min Algorithm

In min-min algorithm, the task which has minimum execution time is assigned to the resource that can process faster. In the proposed approach, constrained min-min algorithm is used. Min-min algorithm is executed inside the clusters with the available fog nodes. If the cluster is busy, the controller node checks for the nearby cluster that has idle fog nodes. The cluster will forward the task with optimum latency. In case the cluster with idle fog node lies far, then the task gets directly forwarded to cloud for processing. It is better to process the task at the cloud datacenter located far away, instead of making the task to wait for long time for preprocessing at fog nodes or making the task to travel longer just for preprocessing.

In case, if two or more idle neighbor nodes are available, then the node with minimum latency is considered for forwarding the task. For calculating latency, two factors are considered: one is number of waiting requests to be served in the clusters and the other is the distance of the idle node from the source node. The minimum distance from source (user) to fog node or cloud datacenter is calculated as given in Eq. (1). It takes the route which has minimum latency, *N,* considering *s* the source from where the task is forwarded, *c* the nearest cloud datacenter, and *n* the number of fog nodes.

$$N = \min\left[[d(s, c)], \min \sum_{i=1}^{n} [d(s, n_i)] \right] \tag{1}$$

3.3 Simulation Results

Algorithms can be tested in real cloud environments as well as using simulation tools. In all cases, real cloud environments cannot be affordable. At the same time, simulations have fruitful benefits such as algorithms that can be tested as many

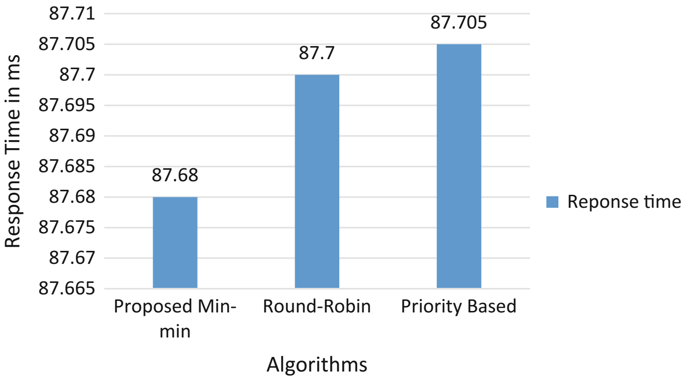


Fig. 2 Comparison of response time with three datacenters

times as required. The proposed algorithm can be compared with other algorithms. Simulations were carried out using Cloud analyst tool.

3.3.1 Configuration Details

The algorithm is tested with the help of Cloud analyst tool. The algorithm is tested based on three different configurations in terms of their response time.

In the first case, three datacenters are created and under each datacenter three fog nodes are configured with three user bases from four different regions. Proposed min-min algorithm is compared with round-robin and priority-based algorithms. The results obtained are compared on the basis of the average response time as shown in Fig. 2. The proposed min-min algorithm obtains response in 87.68 ms, whereas round-robin algorithm obtains response in 87.7 ms and priority-based approach obtains in 87.71 ms.

In the second case, four datacenters are created in four different regions; under each datacenter, four fog nodes are created and five user bases are created from five different regions. The results obtained are depicted as graph in Fig. 3 on the basis of their response time. The proposed min-min algorithm obtains response in 101.35 ms, whereas round-robin algorithm obtains response in 101.37 ms and priority-based algorithm obtains response in 102.63 ms.

In the third case, the five datacenters are created and five fog nodes are configured under each datacenter. In this case, six user bases are created from six different regions. The results obtained are compared in terms of response time which is depicted in Fig. 4. The proposed min-min algorithm obtains response in 101.35 ms, whereas round-robin algorithm obtains in 101.37 ms and priority-based obtains in 102.63 ms.

It is clear from the above results that the constrained min-min algorithm's responses are faster than round-robin and priority-based algorithms. In round–robin

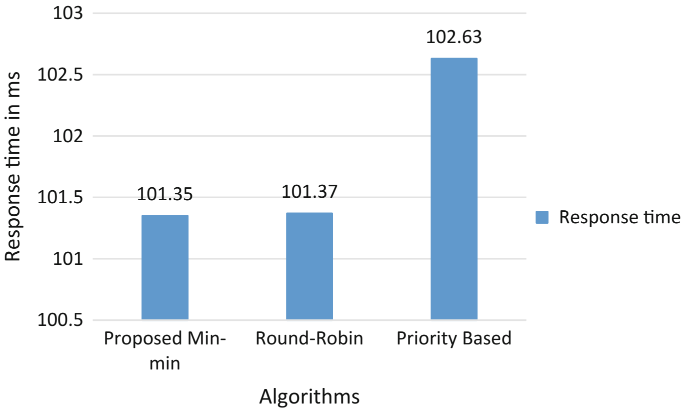


Fig. 3 Comparison of response time with four datacenters

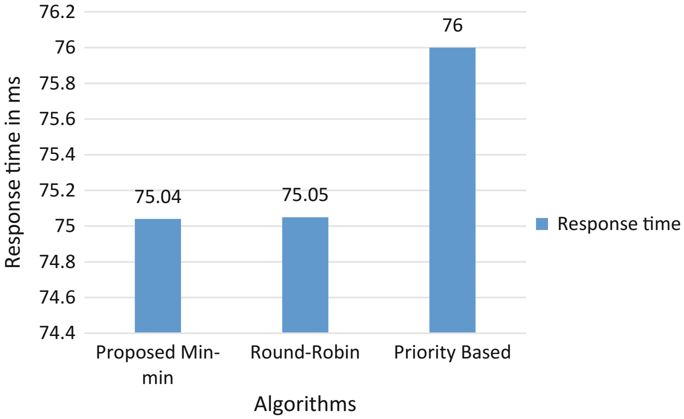


Fig. 4 Comparison of response time with five datacenters

algorithm, the task is evenly distributed among available resources. When execution time of all tasks is equal, this algorithm works well. However, the execution time of all the tasks processing is not equal. Priority-based algorithm sets priority on task and executes the task based on the priority. Setting high priority to prolonging task will make lesser priority tasks to wait longer, whereas the min-min algorithm will sort the task and resources and map the resources minimum execution time to the task with minimum execution time. Instead of checking for idle node availability from large available nodes, it becomes faster when checking in control node metadata alone. Hence, response time decreases. Also, it is better to directly transfer the task to cloud for processing instead of waiting for the idle nodes or transferring along long latency for preprocessing.

4 Conclusion

In the proposed min-min algorithm, the response time obtained is 0.02% better than round-robin and priority-based algorithms. Considering networking resources such as the cluster head, the min-min algorithm has been executed inside each cluster. In case of overloaded cluster, the factors such as distance from the cluster to the neighboring node, the number of waiting tasks to be served in the queue of each cluster, and the distance from the cluster node to the nearest cloud datacenter is considered for forwarding the task. The proposed algorithm has been tested for different numbers of nodes inside the clusters, and the efficiency of the algorithm holds good for smaller clusters nodes. This is the centralized load balancing algorithm that has been implemented using the cloud analyst tool. As a future extension, the proposed algorithm can be implemented with real cloud environment.

References

1. Chamola, V., Tham, C.K., Chalapathi, G.S.S.: Latency aware mobile task assignment and load balancing for edge cloudlets. 2017 IEEE Int. Conf. Pervasive Comput. Commun. Work. PerCom Work. **2017**, 587–592 (2017)
2. Chiang, M., Zhang, T.: Fog and IoT: an overview of research opportunities. IEEE Internet Things J. **3**, 854–864 (2016)
3. Bonomi, F., Milito, R., Zhu, J., Addepalli, S.: Fog computing and its role in the internet of things. In: Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing, pp. 13–16 (2012)
4. Mao, Y., You, C., Zhang, J., Huang, K., Letaief, K.B.: A Survey on Mobile Edge Computing: The Communication Perspective, pp. 1–37 (2017)
5. Al-Shuwaili, A., Simeone, O.: Energy-efficient resource allocation for mobile edge computing-based augmented reality applications. IEEE Wirel. Commun. Lett. **6**, 398–401 (2017)
6. Roy, S., Banerjee, S., Chowdhury, K.R., Biswas, U.: Development and analysis of a three phase cloudlet allocation algorithm. J. King Saud Univ. Comput. Inf. Sci. **29** (2016)
7. Le Tan, C.N., Klein, C., Elmroth, E.: Location-aware load prediction in edge data centers. In: 2nd International Conference on Fog Mobile Edge Computing FMEC, pp. 25–31 (2017)
8. Arya, D., Dave, M.: Advanced informatics for computing. Research **712**, 84–93 (2017)
9. Agarwal, S., Yadav, S., Yadav, A.: An efficient architecture and algorithm for resource provisioning in fog computing. Int. J. (2016)
10. Yu, L., Jiang, T., Zou, Y.: Fog-assisted operational cost reduction for cloud data centers. IEEE Access. **5** (2017)
11. Liu, C.-F., Bennis, M., Poor, H.V.: Latency and Reliability-Aware Task Offloading and Resource Allocation for Mobile Edge Computing (2017)
12. Banerjee, S., Adhikari, M., Kar, S., Biswas, U.: Development and analysis of a new cloudlet allocation strategy for qos improvement in cloud. Arab. J. Sci. Eng. **40**, 1409–1425 (2015)
13. Chen, H., Wang, F., Helian, N, Akanmu, G.: User-priority guided Min-Min scheduling algorithm for load balancing in cloud computing. ieeexplore.ieee.org (2013)
14. Rajput, S.S., Kushwah, V.S.: A genetic based improved load balanced min-min task scheduling algorithm for load balancing in cloud computing. In: 2016 8th International Conference on Computational Intelligence and Communication Networks, pp. 677–681 (2016)
15. Wu, M.-Y., Shu, W., Zhang, H.: Segmented min-min: a static mapping algorithm for meta-tasks on heterogeneous computing systems. In: Proceedings on 9th Heterogeneous Computing Workshop HCW 2000 Cat NoPR00556, pp. 375–385 (2000)

An Empirical Comparison of Six Supervised Machine Learning Techniques on Spark Platform for Health Big Data

Gayathri Nagarajan and L. D. Dhinesh Babu

Abstract Health care is one of the prominent industries that generate voluminous data, thereby finding the need for machine learning techniques with big data solutions. The goal of this paper is to (i) compare the performance of the six different machine learning techniques in spark platform specifically for health big data and (ii) discuss the results from the experiments conducted on datasets of different characteristics, thereby drawing inferences and conclusion. Five benchmark health datasets are considered for experimentation. The metric chosen for comparison is the accuracy, and the computational time of the algorithms and the experiments are conducted with different proportions of training data. The experimental results show that the logistic regression and random forests perform well in terms of accuracy and naive Bayes outperforms other techniques in terms of computational time for health big datasets.

Keywords Health big data · Machine learning · Spark

1 Introduction

With the advent of mobile applications, cloud computing, social media, sensor devices, etc., every individual has become a producer of data apart from being a consumer. Billions of data are generated on a daily basis. Hence, extracting useful insights from this big data and providing a descriptive, predictive, and prescriptive analytics involve lot of challenges. Big data solutions are found to be effective in processing, storage, and handling these data efficiently and quickly. Machine learning techniques are proved to be effective to extract insights from this data. Hence, machine learning techniques on big data solutions play a major role for big data analytics. Though there is lot of machine learning techniques, different techniques are

G. Nagarajan (✉) · L. D. Dhinesh Babu
School of Information Technology and Engineering,
VIT University, Vellore 632014, Tamil Nadu, India
e-mail: gayunagarajan1083@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_32

proved to be effective for different problems depending on the characteristics of the data and other criteria. An empirical comparison of ten different supervised learning techniques on eleven different binary classification datasets is carried out on the basis of eight evaluation metrics in [4]. Different techniques outperform each other with respect to different metrics. Our work emphasizes on health big data since health care is an important sector generating lot of data every second. Sensor devices, wearable devices on patients, intensive care unit monitoring, electronic health reports of patients, doctor's prescriptions, and scanned images are all few sources of big data in healthcare industry. The industry generates heterogeneous data that includes structured, semi-structured, and unstructured data. Extracting hidden knowledge in them to predict the future consequences is of great importance in healthcare industry as they are directly related to the life of the patients. The scope for data mining applications in health care is presented in [6]. Data mining in health care includes various applications like earlier prediction of the diseases, predicting the readmissions rate in the hospitals, decision support system for physicians, etc. They also involve lot of challenges like data unavailability, data integration, data governance, etc. Handling these voluminous data is yet another challenge that requires big data solutions for efficient processing. The need for big data solutions in healthcare industry is presented along with their challenges in [8]. There is always a need to compare the performance of different machine learning techniques on big data platforms for healthcare data which is the motivation behind our work. A comparative analysis on the performance of five different machine learning algorithms for practical IP traffic flow classification is presented in [15]. Though there are many papers in healthcare sector addressing solutions for specific problems, very few papers involve a generalized comparison of machine learning techniques on different health datasets. Hence, our work presents an empirical comparison on the performance of six different machine learning techniques on spark platform in terms of both accuracy and computational time. The paper is organized as follows. Section 2 presents the background knowledge required to understand the paper, Sect. 3 presents the experimental framework, and Sect. 4 discusses the results and draws inferences. Section 5 ends with conclusion.

2 Background Knowledge

2.1 Supervised Machine Learning Techniques

Supervised machine learning techniques are used for classification or regression problems where the class labels are defined. This subsection provides an overview of the six supervised machine learning techniques used in this paper.

Naive Bayes. Naive Bayes is a probabilistic classifier which is based on the concept of Bayes algorithm and works with independent assumptions. Naive Bayes derives the conditional probability for the relationship between an attribute value and a class. The classifier then provides estimation for the probability that a feature is having a certain value [15]. The conditional probability is given by Eq. 1.

$$P(X/Y) = \frac{P(X)P(Y/X)}{P(Y)} \tag{1}$$

where $P(X/Y)$ is the probability of X given Y called posterior probability. $P(Y/X)$ is the probability of Y given X is true. $P(X)$ is the probability that X is true called prior probability. $P(Y)$ is the probability of Y regardless of X. Following the calculation of posterior probability, the maximum probable hypothesis is chosen (Maximum a posteriori MAP hypothesis) and is given by Eq. 2.

$$MAP(X) = \max(P(X/Y) \tag{2}$$

There are different types of naive Bayes algorithm like Gaussian naive Bayes, multinomial naive Bayes, Bernoulli naive Bayes, etc. Naive Bayes is a fast and simple algorithm that can be easily trained on small datasets. Yet, the major disadvantage of naive Bayes is that it cannot learn the relationship between the features [11].

Multilayer Perceptron. Multilayer perceptron is feedforward neural networks containing three layers of nodes: the input, hidden, and the output layers. Multilayer perceptron is used widely as supervised learning technique for classification and regression problems. It uses activation functions like binary, sigmoidal, etc. The sigmoidal function is a special case of logistic function given in Eq. 3.

$$F(X) = \frac{\mathrm{e}^{x}}{\mathrm{e}^{x} + 1} \tag{3}$$

A weighted sum of the input nodes is passed on to the hidden layer and the input from the hidden layer is passed on to the output layers. The activation function is applied at each layer, and the network is trained using the training algorithm. The most commonly used algorithm for training the network is backpropagation. Multilayer perceptron is good at learning nonlinear models and real-time models. Yet, its major disadvantage is to tune a number of parameters like number of nodes, number of hidden layers, weights, iterations, etc.

Logistic Regression. Logistic regression is used for problems where the outcome is a categorical variable, and it describes the linear relationship between the log odds of the outcome and the set of independent explanatory variables [9]. It predicts the probability of the dependent variables based on the independent variables. Logistic regression can be binomial, ordinal, or multinomial depending on the categorical variable. Logistic regression assumes that the dependent variable is a probabilistic event. A logistic curve relates the independent variable X to the rolling mean of the dependent variable Y [2]. The formula is given in Eq. 4.

$$P = \frac{\mathrm{e}^{a+bx}}{1 + \mathrm{e}^{a+bx}} \tag{4}$$

The coefficients of logistic regression are usually estimated using maximum likelihood estimation technique. Logistic regression is an efficient method that yields less variance and works well with decision boundaries. It is also easy to use, and its parameters are easy to interpret, yet suffer from high bias [5].

Decision Trees. Decision trees are tree-like models where the nodes represent the test on attribute, link represents the path taken based on the outcome of the test, and the leaves represent the decision nodes. They can be easily converted to rules. The nodes are chosen based on the concepts of entropy, information gain, etc. Entropy is calculated using Eq. 5.

$$\mathrm{Ent}(T) = -\sum_{i=1}^{n} p_i \mathrm{log} p_i \tag{5}$$

where p_i represents the probability of each class in the child node resulting from the split in the tree. Decision trees can be classification or regression trees based on the target variable. Decision trees are very simple to implement and are flexible for different scenarios. It can be easily combined with other techniques. Yet, it may be complicated if there are many attributes and if many outcomes are linked. There are different decision tree algorithms like ID3, C4.5, CART, CHAID (Chi-squared automatic interaction detector), MARS, etc. [14].

Random Forests. Random forests is a tree ensemble method that constructs multiple decision trees and provides the mean of the predictions or the majority of the predictions as the resulting output. It finds its use for classification and regression tasks. It works on the concept of bagging technique. The final class label for x_1 is given by Eq. 6.

$$F(X_1) = \frac{1}{n} \sum_{n=0}^{m} f_n(x_1) \tag{6}$$

Random forests is expected to yield accurate results compared to other methods since it works by constructing multiple trees by randomly selecting the splits from the best splits. Bagging technique reduces variance without increase in the bias. The major advantage of random forests is its improved accuracy. It works efficiently for big data with many variables or features since it constructs multiple trees in a parallel fashion. The major disadvantage of random forests is that it is found to overfit when the tasks involve noise.

Gradient Boosted Trees. Gradient boosted trees is an ensemble technique that uses the concept of boosting method. It is considered as an optimization function to minimize the loss of the model adding weak decision trees using gradient procedure [3]. Many weak decision trees with leaf nodes are combined together to avoid the problem of overfitting. It is a robust and scalable method [13]. But gradient boosted trees are computationally expensive and require the use of feature selection algorithms to reduce the computational time [1].

2.2 Spark

Spark is an open-source framework for big data processing. It is useful for cluster computing. The application programming interface of spark is built based on resilient distributed database (RDD) that helps to distribute the read-only data on different clusters with fault-tolerant capability [12]. MLlib is the distributed machine learning library of spark. Data reuse is quite common in few machine learning algorithms like logistic regression and k-means clustering as they are iterative in nature and use optimization techniques, and hence RDD works efficiently on them. Spark is proved to be faster than Hadoop [16]. A work was carried out to predict the year of the song using spark platform's MLlib. Random forest yields better accuracy but linear regression outperforms in terms of computational performance. With multiple nodes, the computational time reduced to a greater extent [10].

2.3 Criteria Used for Comparison

There are several criteria used for comparing the different machine learning techniques. Few of them include accuracy, precision, recall rate, computational complexity, etc. Our paper concentrates on two major criteria accuracy and computational performance. Accuracy is the percentage of correct predictions out of all the predictions made and is calculated using Eq. 7.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions made}} \tag{7}$$

Computational time here refers to the total time required for the algorithm to yield the results. It is the time taken to build the model which includes the training time and the time taken to predict for the test data.

3 Experimental Framework

This section gives a description about the health datasets used for experimentation and the results obtained with different machine learning algorithms.

3.1 Dataset Description

Five health datasets are used for experimentation. The datasets are downloaded from Kaggle and UCI machine learning data repository. All the five datasets are examples for binary supervised classification tasks but with different number of observations and features. The nature and the characteristics of the features in each dataset are different. Diabetes dataset (from Kaggle) has 768 observations with eight features. Heart dataset (from UCI machine learning data repository) has 270 observations and 13 features. Spine dataset (from Kaggle) has 310 observations and 12 features. Chronic dataset (from UCI machine learning data repository) has 400 observations with 24 features. Cancer dataset (from Kaggle) contains 569 observations with 31 features.

3.2 Results

The experiments are conducted on the above datasets with six different machine learning algorithms—logistic regression (LR), random forests (RF), decision trees (DT), naive Bayes (NB), multilayer perceptron (MLP), and gradient boosted trees (GBT). Three types of training models are built 70% training data and 30% testing data, 60% training data and 40% testing data, and 50% training data and 50% testing data on each dataset with each machine learning algorithm to analyze their performance with respect to accuracy and computational time. The computational time of the algorithm is computed by executing the algorithms of MLlib of spark on single cluster. The algorithms are expected to yield less computational time with multiple clusters. The results showing accuracy and computational time are given in Tables 1 and 2, respectively.

Table 1 Training model: accuracy

Datasets	Proportion	NB	MLP	LR	DT	RF	GBT
Diabetes	70:30	63.22	74.39	74.79	71.9	**75.62**	71.49
	60:40	65	76.55	**77.53**	71.67	76.88	71.67
	50:50	65.94	72.75	**75.2**	71.66	74.93	69.21
Heart	70:30	80.95	75	**85.71**	77.38	79.77	77.38
	60:40	80.96	80	**87.62**	77.2	80.96	77.2
	50:50	76.61	70.97	**87.9**	68.55	79.03	70.97
Spine	70:30	68.05	80.41	**84.54**	82.47	83.5	81.44
	60:40	68.25	73.02	**85.71**	76.98	76.98	76.98
	50:50	72	76.7	83.33	**85.33**	80	84.67
Chronic disease	70:30	65.11	99.98	99.97	**100**	**100**	**100**
	60:40	64.67	99.98	99.98	**100**	**100**	**100**
	50:50	65.83	99.99	99.99	99.99	**100**	99.99
Cancer	70:30	79.1	99.93	**99.94**	**99.94**	**99.94**	99.93
	60:40	79	99.94	**99.95**	**99.95**	**99.95**	99.93
	50:50	79.93	99.93	**99.95**	99.91	**99.95**	99.91

Bold represent the highest accuracy among all the compared machine learning techniques

Table 2 Training model: computational time in seconds

Datasets	Proportion	NB	MLP	LR	DT	RF	GBT
Diabetes	70:30	2.83	7.70	2.36	**1.60**	1.62	7.86
	60:40	**0.73**	2.74	1.46	0.98	1.11	8.22
	50:50	**0.85**	3.15	1.41	0.93	1.08	8.89
Heart	70:30	**0.75**	3.68	1.75	0.99	1.11	12.27
	60:40	**0.84**	2.92	1.67	1.06	1.18	7.79
	50:50	**0.72**	2.65	1.54	1.01	1.09	8.87
Spine	70:30	**0.70**	2.98	1.95	0.96	1.16	10.43
	60:40	**0.72**	3.76	1.63	1.02	1.08	9.32
	50:50	**0.68**	2.94	2.19	0.95	1.07	8.80
Chronic disease	70:30	**0.99**	1.83	1.98	1.19	1.32	4.22
	60:40	**0.96**	2.46	1.79	1.41	1.28	4.01
	50:50	**0.96**	2.28	1.81	1.18	1.24	5.41
Cancer	70:30	**1.18**	3.03	2.47	1.48	1.60	9.12
	60:40	**1.08**	3.47	2.48	1.40	1.60	9.97
	50:50	**1.36**	4.75	3.67	1.69	1.92	24.35

Bold represent the lowest computational time among all the compared machine learning techniques

4 Discussion of Results and Inferences

The experiments are conducted on five different health datasets with six supervised machine learning algorithms. Three different training models are built for each case with 70:30, 60:40, and 50:50 proportion of training and testing data. Accuracy and computational time are considered as criteria for comparison in our work. The tables above show that random forest and logistic regression outperform other methods in terms of accuracy. There are certain cases where random forest outperforms logistic regression and the reverse. But these two methods are proved to be more accurate in comparison with the other methods. Naive Bayes outperforms other methods in terms of computational time. Though naive Bayes outperforms other methods in terms of computational time, it is poor with respect to accuracy. In fact, its accuracy is the least in comparison with other benchmark methods for diabetes, spine, cancer, and chronic disease datasets. When a comparison is to be drawn between logistic regression and random forests, the computational time for random forests is lesser than logistic regression probably because logistic regression uses iterative procedure like maximum likelihood method. Yet, the other methods multilayer perceptron and gradient boosted trees take more time in comparison with logistic regression. The results prove that random forest is better for supervised learning in health datasets in terms of both accuracy and computational time followed by logistic regression. A work is presented in [9] that compares the logistic regression with gradient boosted trees to predict the landslide susceptibility for multi-occurring landslide events. Both the methods proved to perform equally well in this case. Another work presented in [7] compares three methods: deep neural networks, random forests, and gradient boosted trees. Random forests outperform the other two methods.

5 Conclusion

Our paper presents an empirical comparison of six supervised machine learning algorithms on five health big datasets. Accuracy and computational time are the metrics used for comparing the machine learning algorithms. Three different models are built for each dataset with different proportions of training and testing data. The experiments prove that random forests and logistic regression outperform other methods in terms of accuracy, whereas naive Bayes outperforms other methods in terms of computational time. Yet, the accuracy rate in naive Bayes is very poor in comparison with the other methods used. To resolve the trade-off between random forest and logistic regression, computational time is considered and random forests outperform logistic regression with less computational time.

References

1. Abdunabi, T., Basir, O.: Building diverse and optimized ensembles of gradient boosted trees for high-dimensional data. In: IEEE 3rd International Conference on Cloud Computing and Intelligence Systems (CCIS). IEEE, Piscataway (2014)
2. Brannick, M.T. (n.d.).: Retrieved from http://faculty.cas.usf.edu: http://faculty.cas.usf.edu/mbrannick/regression
3. Brownlee, J.: A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning. Retrieved from https://machinelearningmastery.com: https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/ (2016)
4. Caruana, R., Mizil, A. N.: An empirical comparison of supervised learning algorithms. In: ICML ’06 Proceedings of the 23rd International Conference on Machine Learning, pp. 161–168. ACM, Pennsylvania (2006)
5. Dreiseitl, S., Machado, L.O.: Logistic regression and artificial neural network classification models: a methodology review. J. Biomed. Inform. **35**, 352–359 (2002)
6. Koh, H.C., Tan, G.: Data mining applications in health care. J. Health Care Inform. Manage. **19**, 64–72 (2005)
7. Krauss, C., Do, X.A., Huck, N.: Deep neural networks, gradient-boosted trees, random forests: statistical arbitrage on the S&P 500. Eur. J. Oper. Res. **259**, 689–702 (2017)
8. Kuo, M.H., Sahama, T., Kushniruk, A.W., Borycki, E.M., Grunwell, D.K.: Health big data analytics: current perspectives, challenges and potential solutions . Int. J. Big Data Intell. **1**, 35–38 (2014)
9. Lombardo, L., Cama, M., Conoscenti, C., Marker, M., Rotigliano, E.: Binary logistic regression versus stochastic gradient boosted decision trees in assessing landslide susceptibility for multiple-occurring landslide events: application to the storm event in Messina. Nat. Hazards **79**(2015), 1621–1648 (2009)
10. Mishra, P., Garg, R., Kumar, A., Gupta, A., Kumar, P.: Song year prediction using Apache Spark. In: International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, Jaipur, India (2016)

11. Saxena, R. : Data aspirant. Retrieved from http://dataaspirant.com: http://dataaspirant.com/2017/02/06/naivebayesclassifiermachinelearning/ (2017)
12. Shkapsky, A., Yang, M., Interlandi, M., Chiu, H., Condie, T., Antonio, C.: Big Data Analytics with Datalog Queries on Spark. In: International Conference on Management of Data Proceedings of the 2016, pp. 1135–1149. ACM, San Francisco (2016)
13. Wang, Y., Fen, D., Li, D., Chen, X., Zhao, Y., Niu, X.: A mobile recommendation system based on logistic regression and gradient boosting decision trees. International Joint Conference on Neural Networks. IEEE (2016)
14. Wikipedia: Retrieved from https://en.wikipedia.org/: https://en.wikipedia.org/wiki/Decisiontreelearning (2017)
15. Williams, N., Zander, S., Armitage, G.: A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification. ACM SIGCOMM Comp. Commun. Rev. **36**, 5–16 (2006)
16. Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., et al. (n.d.): Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing. In: NSDI'12 Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation, p. 2. ACM, CA

Novel Text Preprocessing Framework for Sentiment Analysis

C. S. Pavan Kumar and L. D. Dhinesh Babu

Abstract Aim of this article is to propose a text preprocessing model for sentiment analysis (SA) over twitter posts with the help of Natural Language processing (NLP) techniques. Discussions and investments on health-related chatter in social media keep on increasing day by day. Capturing the actual intention of the tweeps (twitter users) is challenging. Twitter posts consist of Text. It needs to be cleaned before analyzing and we should reduce the dimensionality problem and execution time. Text preprocessing plays an important role in analyzing health-related tweets. We gained 5.4% more accurate results after performing text preprocessing and overall accuracy of 84.85% after classifying the tweets using LASSO approach.

1 Introduction

Application of Natural Language Processing in sciences and humanities has become a critical research topic in recent years. Medical domain experts and investigators have been working beyond information present in published reports with new techniques in noisy-text preprocessing. Health Language Processing (HLP) is one of the recent research topics of NLP. Due to the massive availability of information in social media (Twitter, Facebook, RSSFeed, blogs, user forums, message boards), it becomes one of the significant resources for unstructured HLP data over recent years. According to the recent Pew report [1, 2] 67% of the US citizens and two-thirds of world's population in the age group 18–49 are using social media. The number of people using social media continues to hike every year, as considerable number of elderly persons is starting to interact on social media (Table 1).

C. S. Pavan Kumar (✉)
School of Computer Science & Engineering, Vellore Institute of Technology, Vellore, India
e-mail: pavan540.mic@gmail.com

L. D. Dhinesh Babu
School of Information & Technology, Vellore Institute of Technology, Vellore, India
e-mail: lddhineshbabu@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_33

Table 1 Twitter's statistics

Twitter statistics (3rd Jan 2018)	Strength
Monthly unique twitter users	342 M
Tweets from third parties (embedded tweets)	60%
Average tweets per day	58 M
Tweets per second	9100
Languages of tweets available	34[a]

[a] https://dev.twitter.com/web/overview/languages

Table 2 Issues with social text

Issue	Summary
Text cleaning	Removing noise from the text
Tokenization	Classifying section of a string (many techniques)
Lexical error	Identifying grammatical errors
Stemming	Removing derived words
Text clarity	Clear meaning of the sentence (sarcasm)
Tagging	Predicting the meaning of the sequence (text)

The user base of social media is increasing and researchers identified the massive availability of health-related knowledge in the form of text. Many unique opportunities opened up due to social media in patient-oriented Health Care with data-centric NLP methods in unprecedented ways.

Usually, information available on social media consists of misspellings, colloquialisms, HTML tags, scripts, advertisements which makes difficult to detect the actual meaning of the sentence for both humans as well as machines [3, 4]. Allowing those kinds of words in the text increases the dimensionality, and hence classification of text becomes more complicated because each word is considered as an individual dimension. To improve the text classification accuracy and to speed up the classification process we have to remove noise from the text. This process involves two stages: transformations and filtering. While transformations stage consists of removing extra spaces, abbreviation expansion, stemming, removing stop words and finally filtering stage involves feature selection some of them are explained briefly in Table 2.

This article is organized as follows—Sect. 2 Related work, Sect. 3: methodology of text cleaning. Section 4: experimentation and results. Section 5 conclusion.

2 Related Work

Text preprocessing plays a vital role in SA. The accuracy of SVM for sentiment analysis is calculated after preprocessing steps transformations and filtering. Feature selection was carried out using TF-IDF (Term Frequency weighting with Inverse Document frequency) and calculated by Eq. (1).

$$\mathrm{TF} - \mathrm{IDF} = \log(N/\mathrm{DF}) * \mathrm{FF}, \tag{1}$$

where N refers to a number of documents considered, FF refers to Feature Frequency, and DF refers to the number of documents which has a particular term. Used "quadratic programming optimization" along with SVM to test the sentiment analysis accuracy [5, 6]. Jude considered huge dataset of 1.6 Million Tweets from Stanford Twitter corpus and reviews of 400 books over Amazon. Processing such huge number of sentences, it requires lot of time, to reduce the time complexity, they implemented the same framework with the Hadoop which showed better results because of the use of GPGPU programming. Hemalatha proposed a basic framework to reduce the dimensionality of the text by cleaning tweets to find the senti-weighted score and proposed SES algorithm to calculate sentiment over preprocessed twitter data [7]. Few other related works are presented in Table 3.

3 Methodology

Proposed Method for preprocessing consists of four phases. Phase 1 includes extracting required tweets from active Twitter feed server using Twitter API by filtering tweets with no re-tweets as shown in Fig. 1. Phase 2 consists of basic text cleaning: removing of URL's, Hashtags (E.g., #WKKB), Symbols, Punctuations, emoticons, new lines. Phase 3, an important phase which deals with Tokenization, Internet Slang Identification[1], Expanding acronyms (E.g., lol, wtf, etc.,), spell corrector[2] (Norvig spell corrector), Normalization, Stemming/Lemmatization and stop words removal. Finally, the last phase calculates sentiment score of each tweet after preprocessing and before the same.

Phase 4 is Sentiment analysis. Here we are following Dictionary based approach to identify the polarity of the sentence and LASSO[3] for regularization [8].

$$\min\left\{\frac{1}{2}\sum_{i=1}^{N}\left(y_i - \beta_0 - x_i^T\beta\right)^2\right\} \text{subjects to} \sum_{k=1}^{p}|\beta_k| \le t,$$

[1] https://gist.github.com/Zenexer/af4dd767338d6c6ba662

[2] http://norvig.com/spell-correct.html

[3] https://en.wikipedia.org/wiki/Lasso_(statistics)

Table 3 Previous processing techniques

References	Preprocessing method	Dataset	Adv/dis adv	Performance evaluation methods	Sentiment analysis method used	Sentiment analysis accuracy (%)
[5]	Transformation and filtering with chi-squared method	Two datasets (movies, DAT 1400 and DAT2000)	Used SVM which can classify linear and nonlinear problems	Precision, Recall, F-measure	SVM with ten folds	92.87 (FM)
[12]	Expanding Acronym, Removing Numbers & URL's, Removing Negation terms with both prior polarity and *N*-grams	STS-test STS-gold SS-Twitter SE-Twitter SemEval2014	Mentioned many datasets but not mentioned which method to be used for domain	Accuracy and F1-measure For both models	LR, SVM, NB, RF	92.5, 91.3, 90.7, 93.2 (F1 score, resp.)
[13]	Basic cleaning, emoticon, negation, PyEnchant, stemming, stopwords	SemEval 2015 & 2016	Accuracy ↑↑ by identifying apt process for data Stemming with basic method	Optimization of system parameters	Naïve Bayes Multinomial	10–15% ↑ improvement over various combinations and stemming + basic gave best results (80.84%)
[7]	Denoising, slang and URL removal, feature selection	STS-test STS-gold	Used both bi-gram and emoticons as features.	Accuracy	LibLinear	Increased from 84.96 to 85.5%
[10]	Terms standardization, stemming, lemmatization, spell check	Google Play reviews[a] and Movie reviews (IMDB)	Handled internet slangs, abbreviations, and jargons.	Accuracy and F-measure	NB, SVM, ME	82.0331%, 79.6%

[a] https://play.google.com/store?hl=en

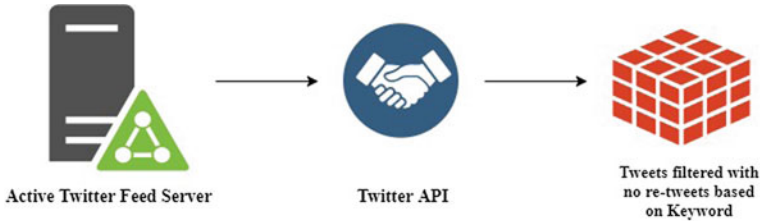

Active Twitter Feed Server
Twitter API
Tweets filtered with no re-tweets based on Keyword

Fig. 1 Phase 1 of preprocessing

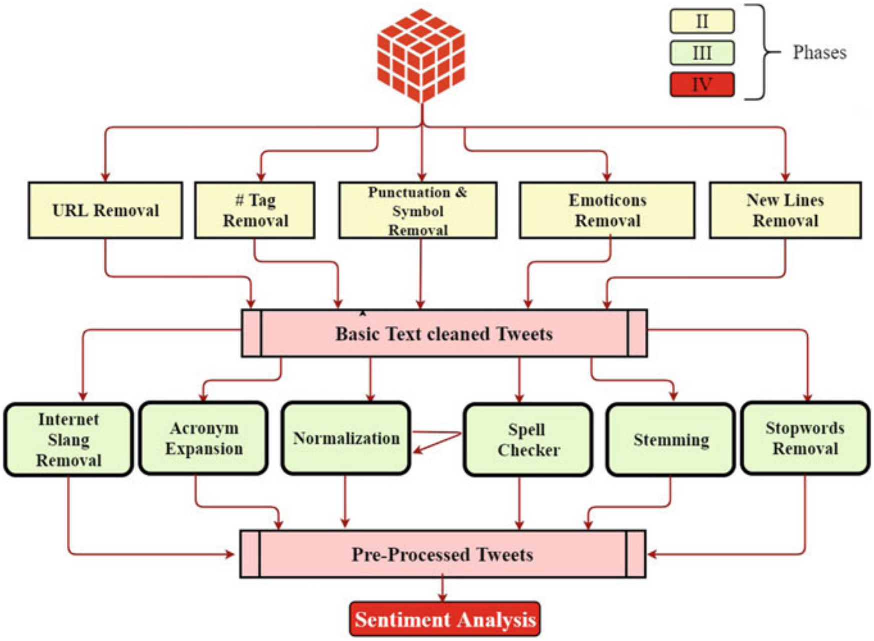
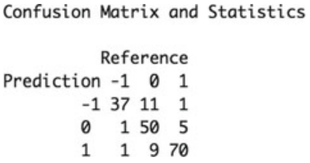


Fig. 2 2, 3, 4 phases of proposed model

where T, is the pre-specified parameter for regression of N cases which has p covariates and a single outcome. The four phases of the proposed model is explained in Figs. 1 and 2.

4 Experimentation and Results

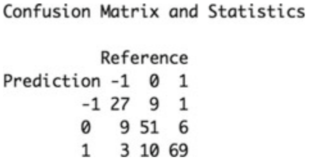
We obtained 84.85% accuracy by following all the four phases mentioned in the previous section. Necessary Text cleaning was done with *R*. Internet slang removal and acronym is done over the internet using Zenexer. Spell checking and normalization



```
Confusion Matrix and Statistics

          Reference
Prediction -1  0  1
        -1 37 11  1
        0   1 50  5
        1   1  9 70
```

Fig. 3 After preprocessing



```
Confusion Matrix and Statistics

          Reference
Prediction -1  0  1
        -1 27  9  1
        0   9 51  6
        1   3 10 69
```

Fig. 4 Before preprocessing

are done using Peter Norvig's algorithm 4. Stopwords removal is done with the help of SMART information retrieval system. Following the completion of phase 3 of the model, LASSO approach has been applied to classifying the tweets and we achieved an accuracy of 84.86% (from the confusion matrix Fig. 3). In our second scenario, we have applied LASSO approach [9] directly over the raw tweets and obtained an overall accuracy of 79.46% (from confusion matrix Fig. 4). We considered unlabeled data for analyzing sentiment over tweets with Adverse Drug Reaction related information and obtained a score for every tweet ranging {−0.5 to 0.5}. After performing normalization, we divided the tweets into three categories based on their scores into Negative (N), Neutral (N), Positive (P) as 1, 0, 1 respectively. The data considered for our experimentation consists of 185 tweets downloaded from the twitter by using the search key @bloodcancer and annotated after reading each and every tweet personally. The data availability in health-related chatter is very less. This is one the major setback in Health Language Processing (HLP).

Results obtained from both the scenarios were compared based on various evaluation metrics. Precision, F1, recall, Detection Rate, Detection Prevalence, Negative Prediction value, Positive Prediction value, Specificity, and Sensitivity were listed in Table 4, which indicates the precision (0.7297) for Class 1 (Negative) of second scenario is decidedly less compared to other classes, and overall accuracy is 79.46%. The precision (0.7551) for Class 1 (Negative) of first scenario is less where the overall accuracy is 86.86%. A Receiver Operating Characteristic (ROC) curve was made between sensitivity and specificity in Figs. 7 and 8 for both the scenarios. The steepness of the curve indicates the efficiency of the LASSO Sentiment Analysis. Most of the authors [3, 7, 10, 11] evaluate the performance of the model via F1, Recall and Precision which was shown between both the scenarios is in Figs. 5 and 6 for two scenarios.

ROC curve allows us to calculate the AUC which helps us to analyze the performance of the classifier. After plotting ROC curves for both the scenarios Multi-class

Table 4 Evaluation metrics

Metric	Before text preprocessing			After text preprocessing		
	Negative (−1)	Neutral (0)	Positive (1)	Negative (−1)	Neutral (0)	Positive (1)
Sensitivity	0.6923	0.7286	0.9079	0.9487	0.7143	0.9211
Specificity	0.9315	0.8696	0.8807	0.9178	0.9478	0.9083
+Pred value	0.7297	0.7727	0.8415	0.7551	0.8929	0.8750
−Pred value	0.9189	0.8403	0.9320	0.9853	0.8450	0.9429
Precision	0.7297	0.7727	0.8415	0.7551	0.8929	0.8750
Recall	0.6923	0.7286	0.9079	0.9487	0.7143	0.9211
F1	0.7105	0.7500	0.8734	0.8409	0.7937	0.8974
Prevalence	0.2108	0.3784	0.4108	0.2108	0.3784	0.4108
Detection rate	0.1459	0.2757	0.3730	0.2000	0.2703	0.3784
Detection prevalence	0.2000	0.3568	0.4432	0.2649	0.3027	0.4324
Balanced accuracy	0.8119	0.7991	0.8943	0.9333	0.8311	0.9147

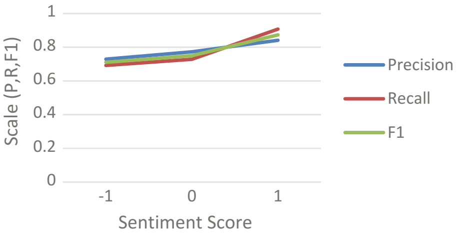


Fig. 5 Precision, recall and F1 score (before preprocessing)

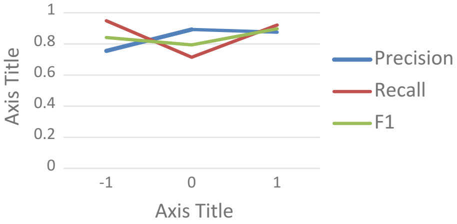


Fig. 6 Precision, recall and F1 score (after preprocessing)

area under the curve before preprocessing and after preprocessing are 86.48 and 92.12% respectively mentioned in Figs. 7 and 8.

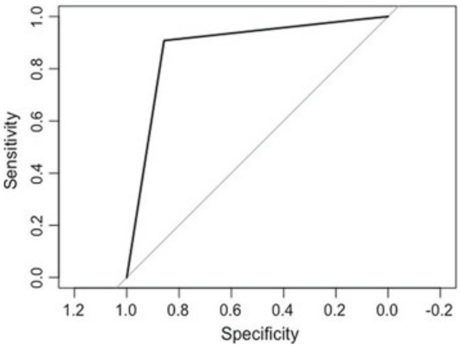


Fig. 7 ROC before preprocessing

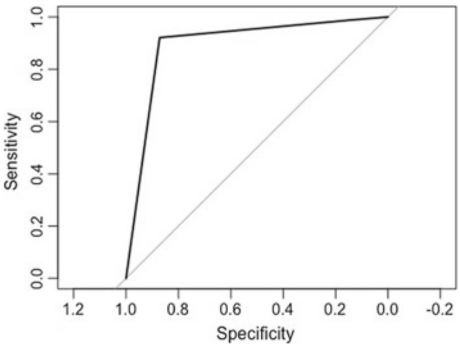


Fig. 8 ROC after preprocessing

5 Conclusion

One of the major challenges in HLP is the availability of the data which can be used by text miner's community to develop new models to explore. Twitter is the easiest way to collect the publicly available data: still we cannot access all the data posted by tweeps because twitter deletes the database from the public once every 6 months. Text preprocessing improves the accuracy of the sentiment analysis because it reduces the dimensionality. The process includes text cleaning, Stopwords removal, stemming, spell check, normalization, emotion removal and slang removal. Significant difference in the accuracy (5.4%) is there between text preprocessed sentiment analysis and non-pre-processed. Dictionaries considered uses English language, where the key term regarding health is present. The accuracy will further increase if we consider lexicon which uses more of medical terminology like Unified Medical Language System (UMLS).

References

1. Fox, S.: The social life of health information. Pew Internet Am. Life Proj, pp. 1–33 (2011)
2. Shearer, E., Gottfried, J.: News Use Across Social Media Platforms (2017)
3. Baldwin, T., Cook, P., Lui, M., Mackinlay, A., Wang, L.: How noisy social media text, how diffrnt social media sources? Proc. IJCNLP, 356–364 (2013)
4. Sarker, A., Gonzalez, G.: Data, tools and resources for mining social media drug chatter. Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining, pp. 99–107 (2016)
5. Angiani, G., Ferrari, L., Fontanini, T., Fornacciari, P., Iotti, E., Magliani, F., Manicardi, S.: A comparison between preprocessing techniques for sentiment analysis in Twitter. In: CEUR Workshop Proceeding 1748 (2016)
6. Nirmal, V.J., Amalarethinam, D.I.G.: Parallel implementation of big data pre-processing algorithms for sentiment analysis of social networking data. Int. J. Fuzzy Math. Arch. **6**, 149–159 (2015)
7. Bao, Y., Quan, C., Wang, L., Ren, F.: The role of pre-processing in twitter sentiment analysis. Lecture Notes in Computuer Science (including Subser. Lecture Notes in Artificial Intelligence, Lecture Notes in Bioinformatics) 8589 LNAI, pp. 615–624 (2014)
8. Prrllochs, N., Feuerriegel, S., Neumann, D.: Generating domain-specific dictionaries using bayesian learning. SSRN Electron. J. (2014)
9. Maurya, A.K., Unit, A.S.: Data Sharing and Resampled LASSO : A Word Based Sentiment Analysis for IMDb Data, pp. 1–19 (2017)
10. Dos Santos, F.L., Ladeira, M.: The role of text pre-processing in opinion mining on a social media language dataset. In: Proceedings—2014 Brazilian Conference on Intelligent System, BRACIS 2014, pp. 50–54 (2014)
11. Hemalatha, I., Varma, D.G.P.S., A. Govardhan, D.: Preprocessing the informal data for efficient sentiment analysis. Int. J. Emerg. Trends Technol. Comput. Sci. **1**, 58 (2012)
12. Haddi, E., Liu, X., Shi, Y.: The role of text pre-processing in sentiment analysis. Procedia Comput. Sci. **17**, 26–32 (2013)
13. Jianqiang, Z., Xiaolin, G.: Comparison research on text pre-processing methods on twitter sentiment analysis. IEEE Access. **5**, 2870–2879 (2017)

Improved Genetic Algorithm for Monitoring of Virtual Machines in Cloud Environment

Sayantani Basu, G. Kannayaram, Somula Ramasubbareddy and C. Venkatasubbaiah

Abstract Resource utilization and energy need to be carefully handled for achieving virtualization in the cloud environment. An important aspect to be considered is that of load balancing, where the workload is distributed so that a particular node does not become overburdened with tasks. Improper load balancing will lead to losses in terms of both memory as well as energy consumption. The resources should be scheduled in a cloud in such a way that users obtain access at any time and with possibly less energy wastage. The proposed algorithm uses an improved Genetic Algorithm that helps reduce overall power consumption as well as performs scheduling of virtual machines so that the nodes are not loaded below or above their capacity. In this case, each chromosome in the population is considered to be a node. Each virtual machine is allocated to a node. The virtual machines on every node correspond to the genes of a chromosome. Crossover and mutation operations have been performed after which optimization techniques have been used to obtain the resulting allocation of tasks. The proposed approach has proved to be competent with respect to earlier approaches in terms of load balancing and resource utilization.

1 Introduction

Cloud computing is a popular computing paradigm. It is an Internet-based environment where the files and applications can be hosted. Thousands of networked

S. Basu (✉) · G. Kannayaram · S. Ramasubbareddy · C. Venkatasubbaiah
School of Computer Science and Engineering, VIT University, Vellore 632 014,
Tamil Nadu, India
e-mail: sayantani.basu2014@vit.ac.in

G. Kannayaram
e-mail: gkannayaram@gmail.com

S. Ramasubbareddy
e-mail: svramasubbareddy1219@gmail.com

C. Venkatasubbaiah
e-mail: cvenkatasubbaiah@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_34

systems can then share and access these resources. Concurrent and distributed computing approaches are used to provide shared resources. Cloud systems generally follow a "pay as you use" model. Customers can use the necessary resources for the required period of time by paying a certain amount.

The set of processing units in cloud are known as virtual machines (VMs) which are allocated to nodes. Nodes should ideally be allotted VMs with the aim of reducing both energy and memory consumption. Otherwise, customer may face issues in scheduling the other resources available. The scheduling algorithm should be designed such that scheduling is carried out effectively with adequate resource utilization. Nodes are capable of being allotted multiple VMs. Care should also be taken that certain nodes are not overloaded and certain nodes are not underloaded. This is handled using the concept of load balancing, where the objective is improving the response time by simultaneously maximizing resource utilization. Load balancing also helps to improve the smooth execution of applications. It is applicable to a variety of scenarios, including heterogeneous and homogeneous environments. Load balancing is predominantly classified into two types: static and dynamic. Static load balancing handles nodes showing low load variation. Dynamic load balancing methods are costlier and more powerful compared to static techniques. They are mostly used in heterogeneous environments. The proposed method also follows dynamic load balancing as the load is distributed with progression of time.

In a cloud computing scenario, if a node is over loaded with a large number of tasks, all the excess VMs should be migrated to underloaded nodes belonging to the same data center. The proposed algorithm also handles this case by using the mutation operator of genetic algorithms. A chromosome (in this case, a node) is mutated or changed if it is overloaded. The process of migrating the tasks takes place and the migrated tasks are allocated to an underloaded node, which is also mutated as a result.

2 Literature Review

Cloud computing is a popular paradigm for delivering services over the Internet that follows a pay-as-you-go model [1]. Cloud providers strive to achieve a healthy balance between operational expenditure and capital expenditure in terms of energy consumption for any cloud system. The cloud environment provides a host of benefits including reliability, robustness and quality of service (QoS). There is a need to strike a balance between the consumption of energy consumption and timely guarantee of resources.

At the same time, managing large amounts of information in data centers with energy efficiency is a challenging task. Several studies have been carried out on energy efficiency in data centers. Energy efficient algorithms have been proposed depending on scheduling of Virtual Machines (VMs) and scaling the speed of the processor [2].

The technique of assigning virtual machines to their respective physical machines is termed as virtual machine placement. VM placement has become a buzzword of late since virtualization is a basic necessity in the paradigm of cloud computing. Proper placement of VMs is crucial in order to improve resource utilization and power efficiency in cloud environment. Several literatures have previously addressed the issue of appropriate VM placement [3, 4].

Chaisiri et al. [5] proposed a linear programming based algorithm for optimal allocation of VMs on physical machines. Their objective was to use the minimum number of nodes. Linear programming formulations have also been considered for server consolidation problems [6, 7]. Van et al. [8] have formulated VM provisioning and VM placement as two constraint satisfaction problems (CSPs). Hermenier et al. [9] have proposed an Entropy resource manager for homogeneous clusters. Their method performs dynamic consolidation based on constraint programming and considers both VM allocation and VM migration.

Goiri et al. [10] have proposed a score-based scheduling that is based on hill-climbing algorithm, to assign every VM onto the physical machine that attains the maximum score. Beloglazov et al. [11] have proposed Modified Best-Fit Decreasing (MBFD), a best-fit decreasing heuristic algorithm, for VM allocation. Their method is power-aware and uses adaptive threshold-based migration. Despite considering energy needs, the time constraints have not been considered. In order to overcome this, Quang-Hung et al. [12] proposed a power-aware genetic algorithm (GAPA) for solving the static virtual machine allocation problem (SVMAP). They have suggested that the computational time needs to be reduced. The deadline of jobs and the migration policies also need to be considered.

The authors in [13] have also proposed a genetic algorithm based approach (GABA) which is capable of adaptively self-reconfiguring the cloud data center VMs having heterogeneous nodes. Their algorithm can also select the physical locations of VMs that are optimal based on time-varying and changing environments. Xu and Fortes [14] have formulated the VM placement problem as a multi-objective optimization problem that aims to minimize the total resource wastage, thermal dissipation, and power consumption costs. They have proposed a modification of genetic algorithm that incorporates fuzzy multi-objective evaluation to search the solution space effectively.

The authors in [15] have obtained a set of non-dominated solutions that minimize both resource wastage and power consumption by applying a multi-objective ant colony algorithm. Feller et al. [16] have proposed a single-objective algorithm based on min-max ant system (MMAS) to minimize the quantity of physical machines necessary to withstand the current load. Gao et al. [17] have focused on both server and network requirements to minimize the energy cost using an energy-aware ant colony algorithm. The authors in [18] have proposed a metaheuristic approach using fireflies that migrates the virtual machine with maximum load to the active node that is least loaded.

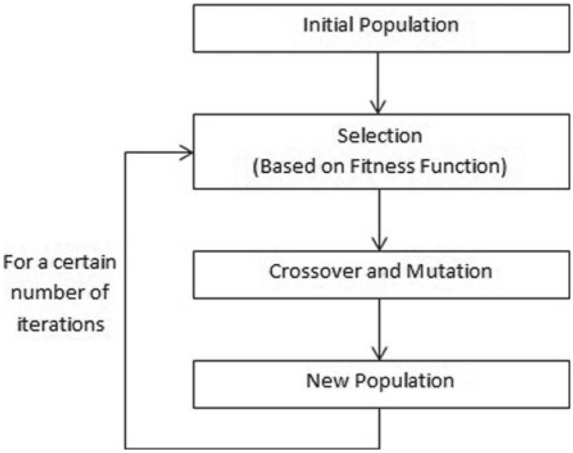


Fig. 1 Flowchart of genetic algorithm

3 Proposed Method

Genetic Algorithms are a class of evolutionary algorithms that are applied to single-objective or multi-objective scenarios. Figure 1 shows the outline of a genetic algorithm. A modified genetic algorithm with local search optimization (GA-LS) has been proposed that is modeled to reduce both the energy consumption and memory usage parameters. This algorithm serves to handle both VM allocation and VM migration. The cloud environment comprises a set of nodes running virtual machines (VMs). The users are presented the virtual environment without providing knowledge of the underlying infrastructures.

3.1 Problem Formulation

The VM allocation problem can be modeled using genetic algorithm and local search terminology with respect to the cloud environment.

3.2 VM Allocation

Genetic Algorithm The cloud environment is considered to be the entire population or candidate solution space. Each node is a chromosome which needs to be optimized according to the fitness function. Individual genes in the chromosome are the VMs allotted. In this case, the fitness function or objective function would be the memory usage and power consumption, both of which need to be reduced. The crossover performed is a single-point crossover where two new individuals are produced.

Equation (1) represents the local fitness function while Eq. (2) represents the global fitness function.

$$ft_{\text{local}}(i, j) = E(i, j) \times M(i, j), \tag{1}$$

where $E(i, j)$ and $M(i, j)$ represent the energy consumption and memory usage of task j to the VMi.

$$ft_{\text{global}} = \sum_{j=0}^{n} ft_{\text{local}} \tag{2}$$

Local Search Local search is a heuristic method applied to search the candidate solution space and locally change or update solutions. In the proposed algorithm, local search is used to select the two best candidates for every crossover operation. In this case, the "best" candidates are those with the best genes, or better fitness in terms of lower energy consumption and lower memory usage.

3.3 VM Migration

If a machine is overloaded with a large number of virtual machines, it is important to mutate the chromosome and migrate the largest gene to the VM that is least loaded in terms of the global fitness function. This process of VM migration helps ensure that the system is not overloaded or underloaded in terms of capacity and ensures adequate load balancing.

4 Implementation and Results

4.1 Dataset

The dataset used for testing the algorithm is the GWA-T-12 Bitbrains which contains data of 1750 virtual machines. The energy and memory parameters were used for the proposed simulation.

4.2 System Configuration

The proposed model (GA-LS) was implemented on a system with 2.4 GHz processor speed and 4 GB RAM.

4.3 Algorithm Simulation

For implementation purposes, Python 3.5 was used for GA-LS. The cloud environment was simulated and the dataset was used to test the algorithm. The algorithm of GA-LS is described as follows:

Proposed Improved Genetic Algorithm GA-LS

```
Start
i = 0;
initialize initial population in cloud for nodes $P_i$;
curr_pop=$P_i$;
evaluate $ft_{local}$ for each gene (virtual machine $VM_j$) in $P_i$;
while number of iterations
        local search selection of  set of fit $P_i$
        perform crossover
        evaluate $P_i$
        if $ft_{global}(P_i) > ft_{global}(P_{i-1})$
        then update curr_pop=$P_i$
        i=i+1;
end_while
set threshold for cloud environment
while number of iterations
        over_loaded={set of chromosomes > threshold}
        under_loaded={set of chromosomes < threshold}
        if $node_i \in$ over_loaded
            mutate chromosome and assign set of  $VM_j$ to node $\in$ under_loaded
end_while
End
```

The proposed algorithm showed competent performance as shown in Fig. 2 for a cloud environment with two nodes considering energy and memory as parameters. It is observed that over a series of epochs, the average fitness and load distribution and resource utilization are carried out evenly by allocating the required number of VMs.

5 Conclusion

In this paper, an improved genetic algorithm with local search (GA-LS) has been proposed that handles load balancing and manages resource utilization. In this case, the objective function aimed at reducing the energy consumption and memory usage of virtual machines allocated to nodes in a cloud environment.

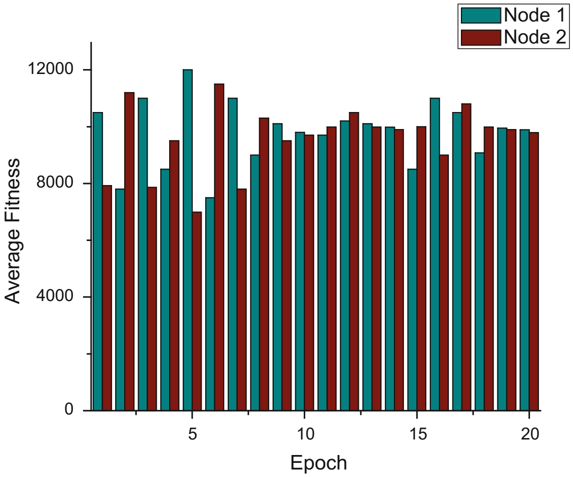


Fig. 2 Average fitness of GA-LS over a series of epochs

The system can be further improved using parallelization and by considering other parameters as objectives for the genetic algorithm. Other approaches can also be included to enhance the performance of the proposed algorithm.

References

1. Zhang, Q., Cheng, L., Boutaba, R.: Cloud computing: state-of-the-art and research challenges. J. Internet Serv. Appl. **1**(1), 7–18 (2010)
2. Albers, S., Fujiwara, H.: Energy-efficient algorithms for flow time minimization. ACM Trans. Algorithms (TALG) **3**(4), 49 (2007)
3. Cardosa, M., Korupolu, M.R., Singh, A.: Shares and utilities based power consolidation in virtualized server environments. In: IM'09. IFIP/IEEE International Symposium on Integrated Network Management, pp. 327–334. IEEE, Piscataway (2009)
4. Grit, L., Irwin, D., Yumerefendi, A., Chase, J.: Virtual machine hosting for networked clusters: Building the foundations for "autonomic" orchestration. In First International Workshop on Virtualization Technology in Distributed Computing VTDC 2006, p. 7. IEEE, Piscataway (2006)
5. Chaisiri, S., Lee, B.S., Niyato, D.: Optimal virtual machine placement across multiple cloud providers. In: Services Computing Conference, 2009. APSCC 2009. IEEE Asia-Pacific, pp. 103–110. IEEE, Piscataway (2009)
6. Bichler, M., Setzer, T. and Speitkamp, B.: Capacity Planning for Virtualized Servers (2006)
7. Speitkamp, B., Bichler, M.: A mathematical programming approach for server consolidation problems in virtualized data centers. IEEE Trans. Serv. Comput. **3**(4), 266–278 (2010)
8. Van, H.N., Tran, F.D., Menaud, J.M.: July. Performance and power management for cloud infrastructures. In: 2010 IEEE 3rd International Conference on Cloud Computing (CLOUD), pp. 329–336. IEEE, Piscataway (2010)
9. Hermenier, F., Lorca, X., Menaud, J.M., Muller, G., Lawall, J.: Entropy: a consolidation manager for clusters. In: Proceedings of the 2009 ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments, pp. 41–50. ACM, New York City (2009)

10. Goiri, I., Julia, F., Nou, R., Berral, J.L., Guitart, J., Torres, J.: Energy-aware scheduling in virtualized datacenters. In: 2010 IEEE International Conference on Cluster Computing (CLUSTER), (pp. 58–67). IEEE, Piscataway (2010)
11. Beloglazov, A., Abawajy, J., Buyya, R.: Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. Future Gener. Comput. Syst. **28**(5), 755–768 (2012)
12. Quang-Hung, N., Nien, P.D., Nam, N.H., Tuong, N.H., Thoai, N.: A genetic algorithm for power-aware virtual machine allocation in private cloud. In: Information and Communication Technology-EurAsia Conference, pp. 183–191. Springer, Berlin, Heidelberg (2013)
13. Mi, H., Wang, H., Yin, G., Zhou, Y., Shi, D., Yuan, L.: Online self-reconfiguration with performance guarantee for energy-efficient large-scale cloud computing data centers. In: 2010 IEEE International Conference on Services Computing (SCC), pp. 514–521. IEEE, Piscataway (2010)
14. Xu, J., Fortes, J.A.: Multi-objective virtual machine placement in virtualized data center environments. In: Proceedings of the 2010 IEEE/ACM Int'l Conference on Green Computing and Communications & Int'l Conference on Cyber, Physical and Social Computing, pp. 179–188. IEEE Computer Society (2010)
15. Gao, Y., Guan, H., Qi, Z., Hou, Y., Liu, L.: A multi-objective ant colony system algorithm for virtual machine placement in cloud computing. J. Comput. Syst. Sci. **79**(8), 1230–1242 (2013)
16. Feller, E., Rilling, L., Morin, C.: Energy-aware ant colony based workload placement in clouds. In: Proceedings of the 2011 IEEE/ACM 12th International Conference on Grid Computing, pp. 26–33. IEEE Computer Society (2011)
17. Gao, C., Wang, H., Zhai, L., Gao, Y., Yi, S.: An energy-aware ant colony algorithm for network-aware virtual machine placement in cloud computing. In: 2016 IEEE 22nd International Conference on Parallel and Distributed Systems (ICPADS), pp. 669–676. IEEE (2016)
18. Kansal, N.J., Chana, I.: Energy-aware virtual machine migration for cloud computing-a firefly optimization approach. J. Grid Comput. **14**(2), 327–345 (2016)

Technology Convergence for Smart Transportation and Emergency Services in Health care

G. Kannayaram and Shubham Saraff

Abstract Smart Transportation is a model to provide multidimensional advantages such as efficient resource allocation, real-time monitoring of traffic, automatic payments and remote patrolling of highways. Many more advantages of Smart Transportation can be derived due to the convergence of technology used in Internet of Things. Using IoT modules this paper proposes to make intelligent vehicle-roadway communication that can in turn improve transportation efficiency save lives and even positively protect the environment from noise- and air-based hazards. Internationally, the roadways used have remained the same for centuries and has seen no drastic technological change. This technology can pace up the development of the highways and roadways networks by acquiring enormous amount of data that can help analyze the flow of traffic. This Big Data enabled application can also act immediately in the favor of authorities in the form of e-ticketing leading to safer travel. IoT enabled roadways can lead to safer travel along with saving time in critical scenarios such as health-related emergencies. All of these combined have unseen possibilities to make this sector much quicker, safer, and efficient with the use of IoT. This paper also proposes two architectures that can be implemented in emergency services for critical help in healthcare domain.

1 Introduction

Internet of Things (IoT) has become an important technology around the globe to integrate in basic application to produce smarter infrastructure, services and ultimately cities. It has widely converged the technology of communications with computational & analytics. IoT has led to rapid growth in Big Data utilization.

G. Kannayaram · S. Saraff (✉)
School of Computer Science and Engineering, VIT University,
Vellore 632 014, Tamil Nadu, India
e-mail: saraffshubham@gmail.com

G. Kannayaram
e-mail: gkannayaram@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_35

The emergence of new application has led to remote monitoring of data in real time. At times, this data is analyzed within the IoT device and produce immediate desired actions. This makes the IoT application extremely time-efficient and cost-effective. Rise of these applications has also led to modernization of administration work that has in turn positively impact the standard.

In this paper, another application of this technology is being proposed to enhance the transportation mechanisms. Using IoT-based communication modules the modern cities can be transformed into a smarter transportation model. This will enable monitoring of traffic, road status, seamless payment at tollbooths and gas stations, better emergency servicing and improved road safety.

This model flips the conventional mechanisms that were slow, cumbersome, and difficult to monitor. It is the real-time data available in this Smart Transportation system that enables the authorities to remotely monitor and alter the traffic flow. This also reduces traffic jams and leads to more efficient resource allocation. Catering to emergency services also become critical and time utilization becomes much faster with the use of the technology proposed in this paper. Two architecture based model is also discussed in proposed method of this paper.

2 Literature Review

There have been several researches based on how to enable transportation using IoT. These findings have extended from different stages of communication in the vehicle to acquiring information from immediate surroundings. Safer roads can be one the greatest consequence of the data being generated by IoT enabled vehicles.

IoT-enabled cars is not restricted to being connected to the Internet. Modern cars are being optimized in various ways and divided into four stages illustrated in Fig. 1. Each stage has a different motive and goal along with business and technological development. The first stage is connecting to the Internet based on demands of the passengers and driver. These range to navigation and in-car entertainment. The next stage is connection to the remote services that provide data like status of the vehicle and safety checks in real time. Connection to surroundings is considered to be the third stage. This leads to safer driving habits due to support by different modules to prevent collision. In the final stage, the vehicle is fully connected and is capable of communication with other vehicles, immediate surrounding, its passengers and the driver. This concept can be extended to the vehicles that are autonomously driven [1].

IoT has stepped into assistance in vehicle parking represented in Fig. 2. This assistance can be divided into certain steps like detection of free parking space followed by transmission of data to the servers that are centrally located. These servers can then cater to the requests made by the application present in the smart phones which are designed to look for free spaces. The next step of servicing the request is to navigate the driver to the selected spot. This application also restricts access to loading spots and residential parking areas. The different layers are connected and

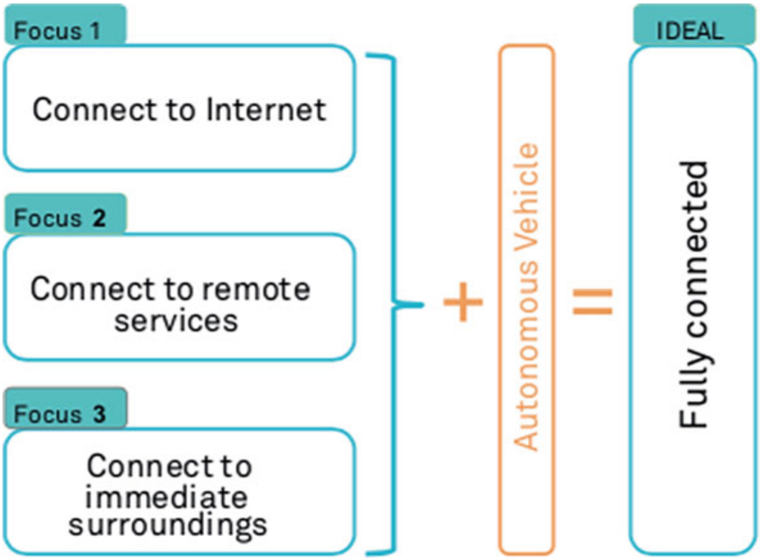


Fig. 1 Different sections of a IoT-enabled car

then executed to function to provide the user with an efficient output using 3G/4G, WSN, and other M2 M techniques [2].

Extending valuable information to the public traveling on road can be of great importance. Information is acquired by using Wireless Sensor Networks (WSN) on the bridges, junctions and highways. This system detects an accident by the intensity of sound produced by it. If the system senses an accident, the information is promptly communicated with the police, hospital, and the other signs preceding the spot of accident. This communication takes place using GSM modem. The network of sensors also detects landslides that occur in hilly areas leading to halt of traffic. Disaster Management authorities can be informed using XBee technologies [3].

The National Traffic Department of Brazil planned to install 90 million vehicles with RFID tag. This is being done in collaboration with Intel. Installing these tags enables the authorities to check status of the car and its licence. Using 915 MHz RFID the cars would be monitored by the authorities. This project is a system to automatically identify the vehicles and stores it in a central and fedrel database [4].

The current Emergency Service Network is outdated and this paper proposes that the use of IoT can enable the homeland security and the emergency services to respond faster. A network of sensors can guide the emergency services like fire engine, police and the ambulance in a much more efficient manner. This can help in cases of great danger like natural disasters, road accidents or even terrorist attacks. ESN is a huge cost to the government and using of IoT can make it efficient and cost effective [1, 5].

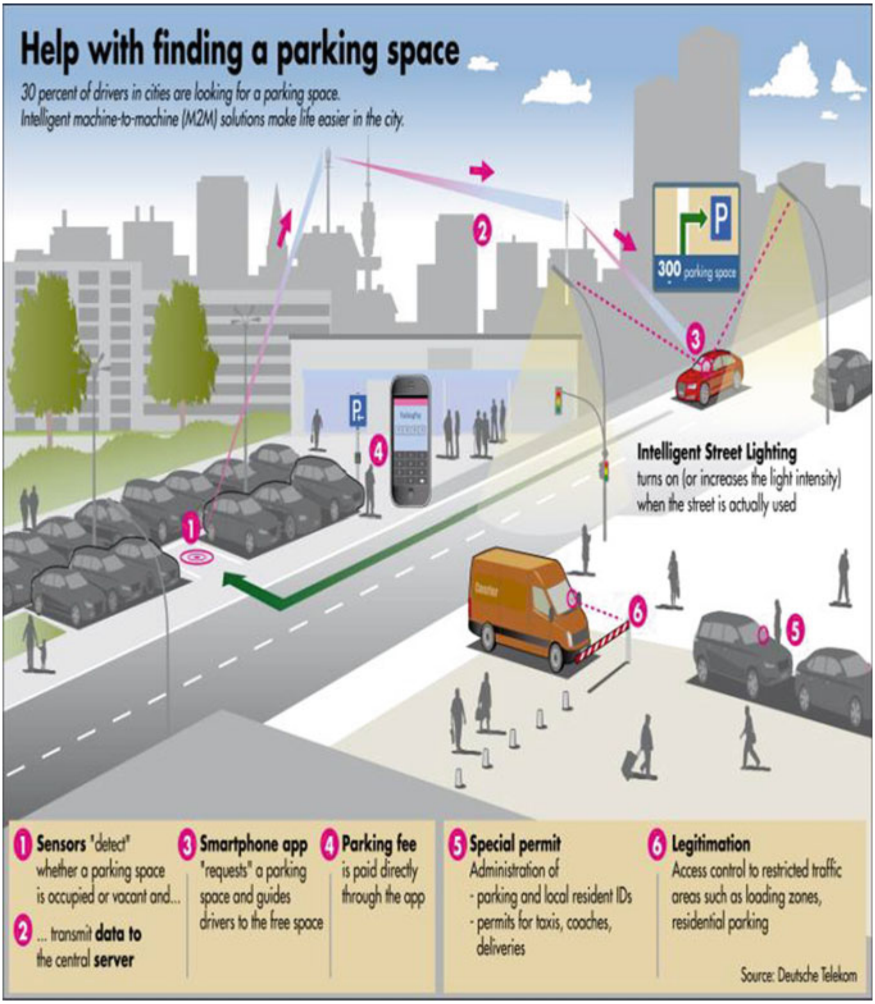

Help with finding a parking space
30 percent of drivers in cities are looking for a parking space.
Intelligent machine-to-machine (M2M) solutions make life easier in the city.
300 parking space
P
Intelligent Street Lighting
turns on (or increases the light intensity) when the street is actually used
1 Sensors 'detect' whether a parking space is occupied or vacant and...
2 ... transmit data to the central server
3 Smartphone app 'requests' a parking space and guides drivers to the free space
4 Parking fee is paid directly through the app
5 Special permit Administration of
- parking and local resident IDs
- permits for taxis, coaches, deliveries
6 Legitimation Access control to restricted traffic areas such as loading zones, residential parking
Source: Deutsche Telekom

Fig. 2 IoT-based parking space finder using M2 M technology

3 Proposed Method

There are several possibilities to make transportation and logistics much smarter, faster and efficient with the use of IoT. The approach of this paper is not only to enable the vehicles with RFID tags but also uplift the technology used on the roads and highways. By establishing a network of sensors across the roads, a system can be created that can analyze the conditions and provide crucial knowledge.

The task of building Smarter roadways is divided into two steps. The first task is to install tags in the vehicle and the second task is establishing a network of sensors on the roads for communication with the central server.

The first task has several factors like storing identification data like license number, expiry date, and other information needed by the authorities. It also has to communicate well with the portals attached on the roadways and has to be tamper free. This is achieved by installing RFID tags with frequency of 915 MHz in every vehicle at the time of license issuance/renewals.

Task two is to build highways and roadways which have technologies that can read the data related to every vehicle and then eventually infer valuable knowledge which can be utilized the by the authorities. Set up can include the following things:

1. RFID transponder which has bit memory of 1024 which in non-volatile
2. Data transfer speed base is 128 Kbps
3. Communications to be protected with AES 120 Cryptography
4. Protection from climate is handled by IP5KO.

Using these four factors it is ensured that the setup follows the security protocols of the country. This is done keeping in the confidentiality and security of the information being stored and transmitted through these Road-based setups.

This dual setup of (i) RFID Tags in Vehicles (ii) Portal setups in the roadways will fetch real-time data about traffic information. These portals will calculate the number of cars passing under it. This information can then calculate the number of cars between portals, resulting in traffic information. Incase of excess number of cars or slow moving vehicles it can notify the departments and they can in turn take corrective measures to improve the traffic flow. This can also help identify the traffic flow pattern at different times leading to better administration of the flow. Helping in estimating economic factors like tax per kilometer, average speed of vehicles and number of accidents.

This system is completely encrypted to protect losing any sensitive date stored or transmitted from the vehicle Tag or the equipment installed along the road. If the tag is tried be altered or tempered then the protected information is destroyed. This way the authenticity of the data can be maintained and official information like license number, model, owner details and account details can be stored.

As the data is authentic, it can be also used to monitor vehicle related crime. A license plate number can be added to list, called blacklisted vehicles. If any car crossing the portal matched the list value then police can track down the location and movement of the vehicle. This method can improve the method used for highway patrol as way urban patrolling. Theft of vehicle or robbery can also be clamped down by this method. This list can also be updated in a national database leading universal access.

This can also lead to learning of information about vehicles crossing state borders and entering or exiting the cities. This can help in tolling and monitoring inflow of people as well as cargo. This is described in Fig. 3.

Another advantage could be synchronization of accounts of the owner with the Tag ID for automatic payments at tollbooth fees, annual tax and also fine. Toll both

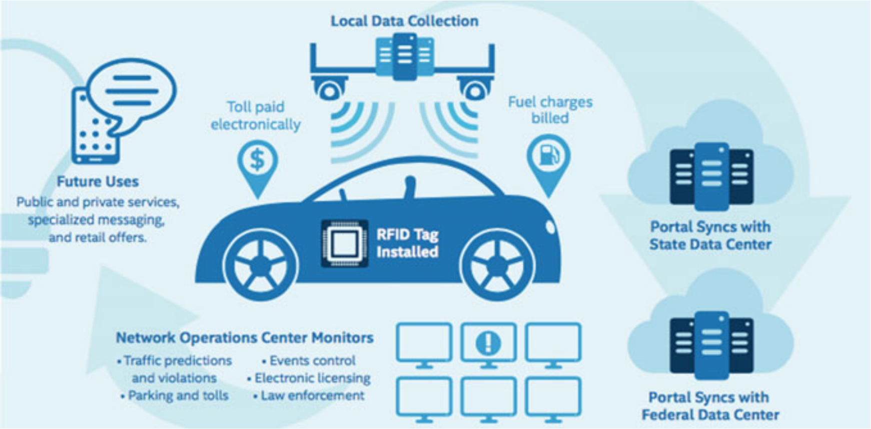

Local Data Collection
Toll paid electronically
Fuel charges billed
Future Uses
Public and private services, specialized messaging, and retail offers.
RFID Tag Installed
Portal Syncs with State Data Center
Network Operations Center Monitors
• Traffic predictions and violations
• Parking and tolls
• Events control
• Electronic licensing
• Law enforcement
Portal Syncs with Federal Data Center

Fig. 3 Graphical explanation of the use of RFID in cars

traffic and infrastructure ad to a huge cost. By integrating this portal systems, the payment process can be made seamless. Greater road safety can be maintained by sending electronic fine to traffic defaulter. Without any need to patrol cars or stop for payment at booths, the average speed at the highways can also safely increase. This paper proposes the use of 915 MHz RFID as the vehicles can pass under the portal at maximum speed of 160 km/h and do not need to slow down for the same.

The mechanism of automatic payments can be extended to payments at petrol pumps or gas stations. After the refueling is complete, the portal placed right above the car deducts the final amount. An e-receipt is generated and sent to the owner's email and registered mobile number.

A two step process is followed in case of emergency services. The architecture of the process is as follows:

1. Mapping the ambulance and the hospital.
2. Real-time update of critical health information of the patient.

Architecture 1 is an efficient model of the Ambulance and Hospital Database which can be stored on the cloud for dynamic updation and use. This can also help in demand and supply model with precision mapping for increased efficiency and faster response time. The Ambulance DB shall contain information related to the geolocation, equipments, contact and status while the Hospital DB shall contain information of the availability, specialty, geolocation, and contact coordinates. All this can be accessed by the user at the device application level with client–server communication. Figure 4 describes Architecture 1.

In emergency services, time plays a critical role and hence information dissemination in real time is critical. Architecture 2 send real-time data of the patient vitals to the doctor or hospital staff for immediate help during the journey to the hospital. This second by second update of Blood Pressure, Heart Rate, Sugar and Electrocardiogram (ECG) can help the hospital guide the ambulance staff better. All the metrics

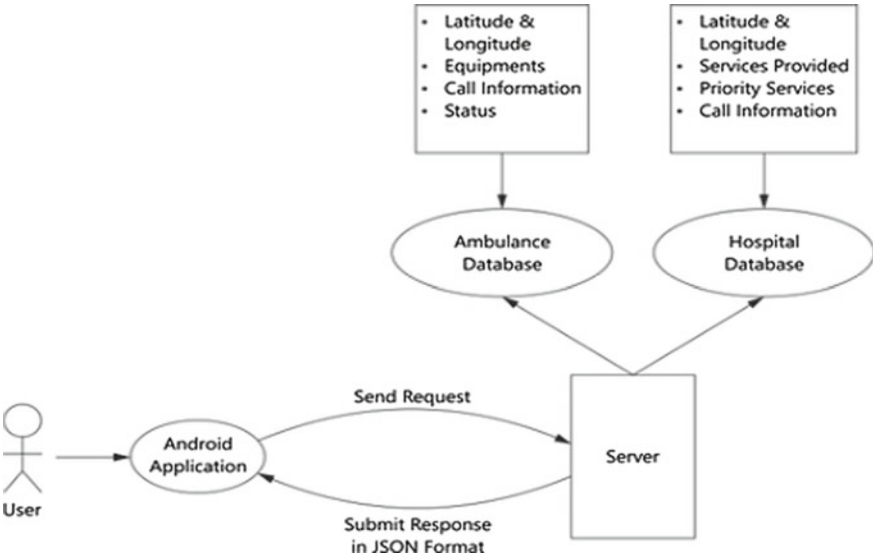


Fig. 4 Architecture 1 mapping the hospital and ambulance

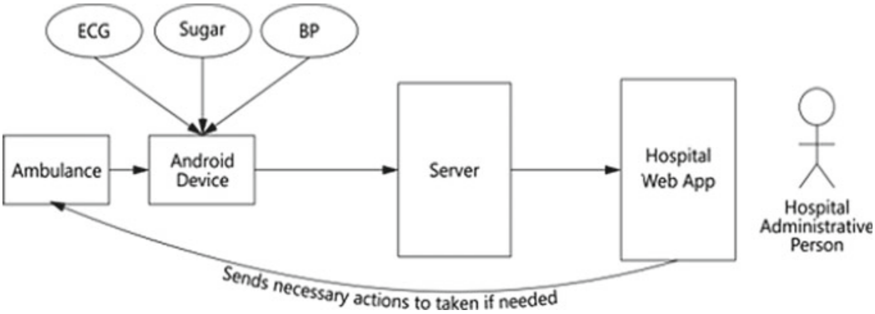

ECG
Sugar
BP
Ambulance
Android Device
Server
Hospital Web App
Hospital Administrative Person
Sends necessary actions to taken if needed

Fig. 5 Real-time data sharing with hospital from ambulance

from the equipments in the ambulance can be streamed live to the hospital saving several minutes of critical time represented in Fig. 5.

The emergency services can also monitor live traffic, fastest, and shortest route to the selected hospital in Architecture 1. This architecture can also select the ambulance based on the arrival time and equipment requirement of the patient.

4 Conclusion

The rapid advent of IoT-based technologies has led to smarter cities and this paper proposed the use of these converging technologies for making smarter infrastructure for transportation. This has been done by analyzing various studies based on installing of RFID technologies of different frequencies in vehicles and how it can impact applications like finding free parking spaces and also respond to emergency services. In this paper, the set up has been described in detail for making smarter roadways which can enable automatic payments, better response systems, and constant monitoring. This is also an attempt to make transportation faster and safer with better time utilization.

References

1. Huawei: Connecting to the Smart Future—Smart Transportation (2016)
2. Sherly, J., Somasundareswari, D.: Internet of Things Based Smart Transportaion System (2015)
3. Adwani, A., Madan, K.H., Hande, R.: Smart Highway Systems for Future Cities (2015)
4. Intel Case study: Brazil Drives into the Future (2015)
5. Redhat: Smart Transportation Applications in the Internet of Things (2016)

A Honey Bee Inspired Cloudlet Selection for Resource Allocation

Ramasubbareddy Somula and R. Sasikala

Abstract The mobile cloud computing has been introduced to address limitations of mobile devices. These limitations include battery lifetime, storage capacity, and processing power. User can offload resource intensive mobile application into cloud for processing can return the result to device. This increases transmission delay and power consumption. The new cloudlet concept has been proposed to address problems such as long delay and huge power consumption. But selection of cloudlet in mobile cloud computing is an important issue for offloading task into cloudlet data center. In this paper, we proposed honey bee inspired cloudlet selection for resource allocation for resource intensive applications from users. The proposed algorithm not only selects low loaded cloudlet but also perform load balancing among cloudlets in such way that the waiting time of user offloading task in queue will be reduced. The experimental result shows that performance of the proposed algorithm optimal compare with existing algorithm. The proposed algorithm represents significant improvement in reduction of user's waiting time in queue.

1 Introduction

Mobile devices become an essential in our lives, because of many advantages such as make a call, organizing meetings, video calls, create different formats of documents and can operate social websites (twitter, Facebook, yahoo, etc.). On the other hand, the cloud computing (CC) is an emerging technology in IT (information technology) industries in order to reduce economical and management cost. Cloud is usually collection of powerful servers which can be interconnected trough network and provide resources to users through Internet as service based on pay-as-you-use model. Even mobile is powerful device to provide all kind of services to user but it has many weakness like battery life, storage capacity, and processing capacity and also stopping user from using mobile application efficient way. The solution for limitations of mobile

R. Somula (✉) · R. Sasikala
VIT University, Vellore, Tamil Nadu, India
e-mail: svramasubbareddy1219@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_36

device is, integration of both mobile computing and cloud computing technologies which is called as mobile cloud computing (MCC). In MCC, the intensive task can be transferred to cloud where enough resource is available to process intensive tasks in less time compare to mobile device and result is returned back to mobile device. With this technique, MCC address both limitations of mobile device such as limited storage and limited processing capacity. Additionally, the large size of images, files, video tapes can be stored at cloud; the user can access stored files whenever required. Even cloud providing enough resources for user offloading applications but still latency makes this approach insufficient. To solve high latency problems, [1].

2 Background

This paper [2] discuss the combination of both cloud computing and mobile computing are involved in deployment of MCC. It is also describes that real-time MCC challenges, scope and growth. This paper developed MobiCloud model at university of Arizona state to simplify the process of MCC. The growth of networked sensors increasing day by day to collect real-time information of different aspects of life including health, military, transport and different augmented reality applications [3]. Therefore, huge amount of data generated through networked sensors that need to be stored in cloud, user can access back whenever required [4]. In this work, the author discussed factors which influence to mobile power consuming when associated with cloud. They presented comparing power consumption between local mobile device and remote cloud by using their own measurements [5]. The author in [6], proposed VM scheduling jobs among cloudlets and handle them inside the cloudlet, and different metrics involved in job allocation: VM overload, allocation of cloudlets to incoming jobs and scheduling VMS. The work in [7], had discussed VM management and deployment in CC using cloudsim toolkit, and concluded that cloudsim tool efficient for reducing power consumption by scheduling VMs.

In this paper [8, 9] Karalooga has proposed the intelligent foraging behavior known as Artificial Bee colony Algorithm (ABC). This algorithm has ability of global search in order to optimize numeric function optimization [10]. The ABC algorithm have been used in different areas to address several problems such as signal processing [11].

3 Cloudlet Selection Inspired Honey Bee Behavior Algorithm

Mobile cloud computing technology has been developed with combination of various technologies such as: mobile computing, cloud computing, and networking. Users can offload their mobile application to cloud to improve battery utilization of mobile devices. But still mobile is suffering with latency and response time problems. The effective load balancing model can be effective by distributing and maintaining load among VMs and also reduce makespan and response time. The completion time can be defined as makespan. We denote completion time of task T_K on VM_l as CT_{kl}. The make can be represented as following function.

3.1 Finding Cloudlet Using Euclidean Distance

$$\text{Cloudlets CL}\{\text{CL}_1, \text{CL}_2, \text{CL}_3, \ldots \text{CL}_n\} \text{ each cloudlet has } R \text{ and } D \qquad (1)$$

$C_i = [R_i, D_i] R_i =$ resource of ith cloudlet, $D_i = [x_i, y_i]$ geographical co-ordinates.

Find Euclidean distance between origin of the request and the cloudlets.

Let the origin of the request be C_0: $C_0 = [R_0, D_0]$ and $D_0 = [x_0, y_0]$

$$\text{Euclidean distance } d_i \Rightarrow \sqrt{(y_0 - y_i) + (x_0 - x_i)} \qquad (2)$$

$D = \{d_1, d_2, d_3, \ldots, d_n\}$ Threshold distance $= t_d$ (assuming that threshold distance).

```
While job do
        SearchCloudlet (job)
                              $if\ R_o < R_i \&\&\ d_i \leq t_d$
      Initialize Load Balancing Mechanism;
        Else
   i++; (search for another cloudlet)
        End if
End while
```

$$\text{Make span} = \max\{\text{CT}_{kl} | k \in T, \quad k = 1, 2 \ldots n \text{ and } l \in \text{VM}, \quad l = 1, 2 \ldots m\} \qquad (3)$$

Response time can be defined as the amount of time taken for both transmission time of request to cloudlet and the response that is produced from cloudlet.

Set of virtual machines and tasks Let $\mathrm{VM} = \{\mathrm{VM}_1, \mathrm{VM2} \ldots \mathrm{VM}_m\}$ and $T = \{T_1, T_2, \ldots T_n\}$. We denote non preemptive tasks as npt. The goal of this model is reduce makespan $\mathrm{CT}_{\max}$ of task T_k finishing time by CT_k. The processing time of task T_k on virtual machine VM_l denoted as P_{kl}.

$$P_l = \sum_{l=1}^{n} P_{kl} \quad l = 1, 2, \ldots, m \qquad (4)$$

We can get Eq. (3) by minimizing $\mathrm{CT}_{\max}$ and Eq. (4) from both Eqs. (2) and (3)

$$\sum_{k=1} P_{kl} \leq \mathrm{CT}_{\max} l = 1, \ldots, m \tag{5}$$

$$\Rightarrow P_j \leq \mathrm{CT}_{\max} \quad l = 1, 2, \ldots, m \tag{6}$$

The finishing time of each task may vary due to load of each VM.

$$\mathrm{CT}_{\max} = \left\{ \max_{k=1}^{n} \mathrm{CT}_k \max_{l=1}^{m} \sum_{k=1}^{n} P_{kl} \right\} \tag{7}$$

The capacity of cloudlet can be defined as follows

$$\mathrm{cap}_l = \mathrm{pe}_{\mathrm{numl}} \times \mathrm{pe}_{\mathrm{mipsl}} + \mathrm{vm}_{\mathrm{bwj}} \tag{8}$$

Here $\mathrm{pe}_{\mathrm{numl}}$ represents number of processing elements, pe $_{\mathrm{mipsl}}$ represents million instruction per second and $\mathrm{vm}_{\mathrm{bwj}}$ bandwidth of VM.

The capacity of all VMs

$$c = \sum_{l=1}^{m} c_j \tag{9}$$

The capacity of all VMs can be treated as capacity of cloudlet.

The length of the task can be represented as load on VM

$$L_{V,M_i,t} = \frac{N(T, t)}{s(\mathrm{VM}_i, t)} \tag{10}$$

The load of all VMs calculated as follows:

$$L = \sum_{l=1}^{m} L_{\mathrm{VM}_l} \tag{11}$$

Each VM processing time can be calculated as follows:

$$\mathrm{PT}_l = \frac{L_{\mathrm{VM}_l}}{C_l} \tag{12}$$

All VMs processing time can be calculated as follows:

$$\mathrm{PT} = \frac{L}{C} \tag{13}$$

Load of Standard deviation:

$$\sigma = \sqrt{\frac{1}{m}\sum_{l=1}^{m}(\mathrm{PT}_i - \mathrm{PT})^2} \tag{14}$$

If the standard deviation (σ) of VMs load is less than given threshold condition (Tc) then system is balanced.

$$\begin{gathered} \text{If}\, \sigma \leq \text{Tc} \\ \text{System is balanced} \\ \text{Exit} \end{gathered}$$

The load of the current VM group exceeds capacity of this group, and then hat load balancing that group is overloaded. In this case, load balancing model is not applicable.

$$\begin{gathered} \text{If}\, L \leq \text{maximum capacity} \\ \text{Not possible} \\ \text{Else} \\ \text{Possible} \end{gathered}$$

all balanced VMs are included into set. Different task priorities are identified by VM as follows

$$T_{\text{high}} \rightarrow \mathrm{VM}_d|\min\left(\sum T_{\text{high}}\right)\epsilon \mathrm{VM}_d \tag{15}$$

$$T_{\text{middle}} \rightarrow \mathrm{VM}_d|\min\left(\sum T_{\text{high}} + T_{\text{midlle}}\right)\epsilon \mathrm{VM}_d \tag{16}$$

$$T_{\text{low}} \rightarrow \mathrm{VM}_d|\min\left(\sum T\right)\epsilon \mathrm{VM}_d \tag{17}$$

In our model, tasks are categorized into three categories (high, middle, low). When high priority task is removed from overloaded VM, it should be placed into idle or normal VM.

3.2 Cloudlet Selection and Load Balancing with Honey Bee's Algorithm

Step1. Find distance and recourse of available cloudlets based on equation (1) and (2), if cloudlet found as per requirements then load balancing algorithm will be applied to among VMs in cloudlet.

```
While job do
    For i=1 to n
        SearchCloudlet (job)
            $if\ R_o < R_i \&\&\ d_i \leq t_d$
                Initialize Load Balancing Mechanism;
            Else
                i++; (search for another cloudlet)
            End if
End while
```

Step2. Find load of all VMs from equation (3),(4),(5) and (6) and also find system is balanced or not:

```
If$\sigma \leq Tc$
    System is balanced
Exit
```

Step3. The decision of load balancing will be made as follows

```
If $L \leq maximum\ capacity$
    Not possible
Else
    Possible
```

Step4. VMs can be formed as a group based on IDLE, NORMAL and OVERLOADED

Step5. Load balancing:

Available recourse in $\mathrm{IDLE_{VM}}$

$$\mathrm{IDLE_{VM_J}} = \text{maximum capacity} - \frac{load}{capacity}$$

Required recourse in $\mathrm{OVERLOADED_{VM_J}}$

$$\mathrm{OVERLOADED_{VM_J}} = \frac{load}{capacity} - \text{maximum capacity}$$

VMS in $\mathrm{IDLE_{VM_J}}$ are sorted in ascending order.

VMs in $\mathrm{OVERLOADED_{VM_J}}$ are sorted in descending order.

While $\mathrm{IDLE_{VM_J}} \neq \varphi$ and $\neq \varphi$

For s=1 to # ($\mathrm{OVERLOADED_{VM}}$) do

Priority tasks are sorted in VM

Each task in queue of VM will find demand machine $\mathrm{VM_d} \in \mathrm{IDLE_{VM}}$

```
            If (T is non preemptive)
        $T_{high} \rightarrow VM_d|min(\sum T_{high}) \epsilon VM_d and Load_{VM_d} \leq capacity_{VM_d}$
        $T_{middle} \rightarrow VM_d|min\left(\sum T_{high} + T_{midlle}\right) \epsilon VM_d and Load_{VM_d} \leq capacity_{VM_d}$
        $T_{low} \rightarrow VM_d|min\left(\sum T\right) \epsilon\ VM_d\ \epsilon\ VM_d and\ Load_{VM_d} \leq capacity_{VM_d}$
    If (T is preemptive)
        $T_{high} \rightarrow VM_d|min(\sum T_{high})\ \epsilon\ VM_d$
        $T_{middle} \rightarrow VM_d|min(\sum T_{high} + T_{midlle})$
        $T_{low} \rightarrow VM_d|min\left(\sum T\right) \epsilon VM_d$
 Update both number of task and priority task assigned to $VM_d$
Update load on both VMs,VMd
Update IDLE, NORMAL and OVERLOADED sets.
        VMS in $IDLE_{VM_J}$ are sorted in ascending order.
        VMs in $OVERLOADED_{VM_J}$ are sorted in descending order.
```

4 Result Analysis

Mobile cloud computing is an emerging technology because integration of multiple technologies. The cloudlet (datacenter) is a small scale datacenter which is distributed across various geographical regions. Users can access cloudlet based on different constraints. In our model, we have considered two parameters for selecting cloudlet: (1) Distance (*D*) (from mobile device to datacenter) and (2) available Resource (*R*) (number of available VMs in datacenter).

Figure 1 describes the makespan with and without Honey Bee Load Balancing. The *X*-axis represents the number of task and *Y*-axis represents the execution time in seconds. Figure 2 states the response time of VMs with different scheduling algorithms. The *X*-axis represents number of tasks and *Y*-axis represents respond time in seconds.

5 Conclusion and Future Work

This paper proposed a conceptual framework which is mainly focusing on selection of distributed resourceful cloudlet in geographical locations with help of load balancing algorithm. Cloudlet mimics remote cloud but it is available around user in order to reduce user request waiting time at queue. In this model, we are involving two

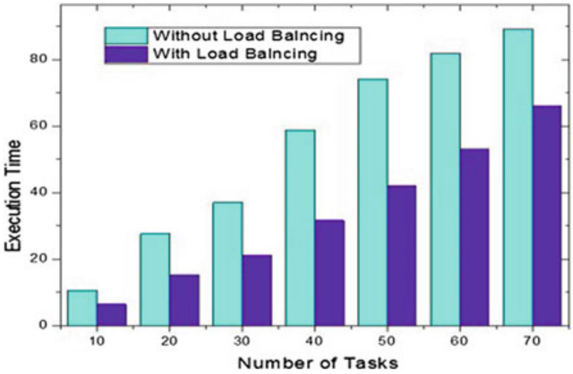


Fig. 1 Comparison of makespan with & without proposed algorithm

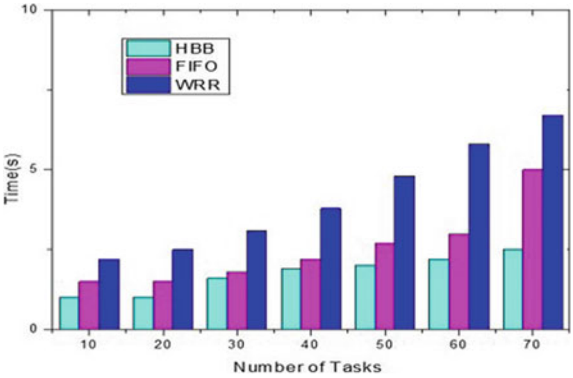


Fig. 2 Response time of VMs in seconds for Honey Bee, FIFO and WRR

parameters to find optimal cloudlet: they are recourse and distance. Once the cloudlet is found according to user service request requirement then Honey Bee load balancing will be initialized for selected cloudlet. These algorithms deals with VMs inside cloudlet as well as consider priorities of incoming task and also remove task from overloaded VM in order to get placed in another underloaded VM. We have compared our proposed algorithm with existing or conventional algorithm. Results show that our algorithm performed well in all cases. Proposed model consider Resource and distance as a main parameters to find optimal cloudlet. In further discuss another factor to improve cloudlet discovery in MCC.

References

1. Satyanarayanan, M.: Mobile computing: the next decade. In: Proceedings of the 1st ACM Workshop on Mobile Cloud Computing & Services: Social Networks and Beyond (MCS) (2010)
2. Huang, D., et al.: Mobile cloud computing. IEEE COMSOC Multimed. Commun. Tech. Comm. E-Lett. **6**(10), 27–31 (2011)

3. Lo'ai, A.T., Bakhader, W.: A mobile cloud system for different useful applications. In: IEEE International Conference on Future Internet of Things and Cloud Workshops (FiCloudW), pp. 295–298 (2016)
4. Lo'ai, A.T., Bakheder, W., Song, H.: A mobile cloud computing model using the cloudlet scheme for big data applications. In: 2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), pp. 73–77 (2016)
5. Miettinen, A.P., Nurminen, J.K.: Energy efficiency of mobile clients in cloud computing. HotCloud **10**, 4 (2010)
6. Shiraz, M., Gani, A.: Mobile cloud computing: critical analysis of application deployment in virtual machines. In: Proceedings of International Conference on Information and Computer Networks (ICICN'12), vol. 27 (2012)
7. Bahwaireth, K., Benkhelifa, E., Jararweh, Y., Tawalbeh, M.A., et al.: Experimental comparison of simulation tools for efficient cloud and mobile cloud computing applications. EURASIP J. Inf. Secur. **2016**(1), 1–14 (2016)
8. Karaboga, D.: An Idea Based on Honey Bee Swarm for Numerical Optimization (2005)
9. Karaboga, D., Basturk, B.: On the performance of artificial bee colony (ABC) algorithm. Appl. Soft Comput. **8**(1), 687–697 (2008)
10. Rao, R.S., Narasimham, S.V.L., Ramalingaraju, M.: Optimization of distribution network configuration for loss reduction using artificial bee colony algorithm. Int. J. Electr. Power Energy Syst. Eng. **1**(2), 116–122 (2008)
11. Singh, A.: An artificial bee colony algorithm for the leaf-constrained minimum spanning tree problem. Appl. Soft Comput. **9**(2), 625–631 (2009)

A New Indexing Technique for Healthcare IOT

P. Sony and N. Sureshkumar

Abstract Healthcare IOT (H-IOT) is now offered as a technology enabler for gathering real-world health-related information. H-IOT index gathers, parses, and stores information to encourage quick and precise data retrieval. Indexing is a useful technique in every search and retrieval system. Indices are utilized to rapidly find information without searching each column in a database each time it is gotten too. The motivation behind to use an index is to upgrade speed and execution in finding applicable health reports for an inquiry question. Without an index, searcher would filter each report in the corpus, which would require impressive time and processing power. Since H-IOT has to accommodate two distinct data—data from medical records and data from IOT-Medical Devices (IOT-MD)—Indexing in H-IOT is not an easy task. Indexing the real-time streaming data is a major barrier to the development of healthcare IOT. Here, we have proposed a new indexing technique based on finite state automata for healthcare IOT. The performances are evaluated and compared with the existing indexing techniques, and the proposed technique outperforms all the existing techniques.

1 Introduction

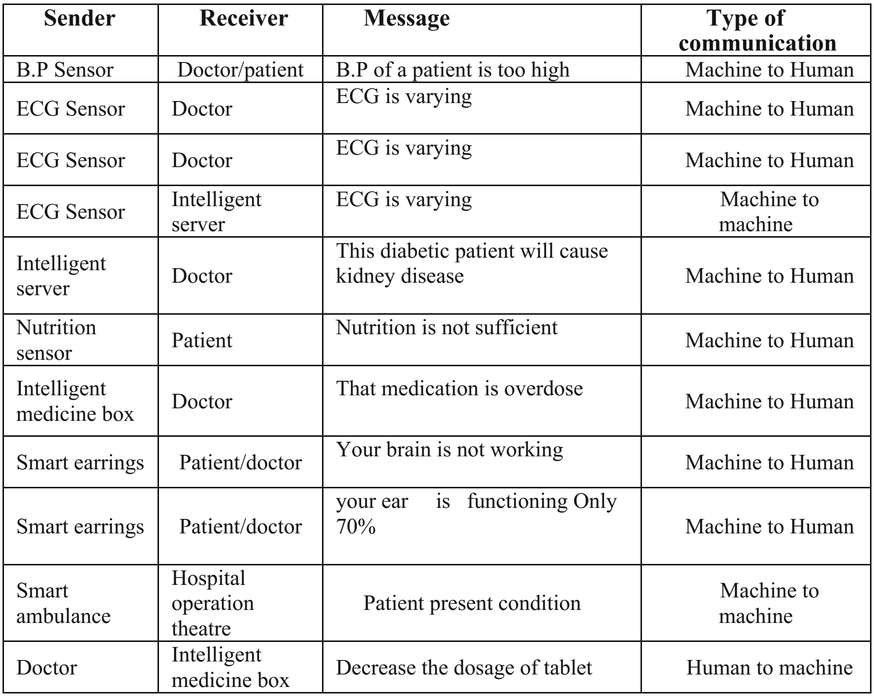
Healthcare IOT (H-IOT) or Internet of Health (IoH) is a system of intelligent articles associated in the e-Health world. These articles can be wearables joined to a patient body, embedded gadgets, sensor associated with the therapeutic hardware, a keen PC, brilliant medicine box or a smart data server, and so forth.

Analysts in the restorative field, therapeutic specialists, community caregivers, and insurance agencies may surf medicinal records for various reasons. Good indexing techniques necessitated making this value-based searching easy. For instance, a

P. Sony (✉) · N. Sureshkumar
Vellore Institute of Technology, Vellore, India
e-mail: spsony@gmail.com

N. Sureshkumar
e-mail: sureshkumar.n@vit.ac.in

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_37



Sender	Receiver	Message	Type of communication
B.P Sensor	Doctor/patient	B.P of a patient is too high	Machine to Human
ECG Sensor	Doctor	ECG is varying	Machine to Human
ECG Sensor	Doctor	ECG is varying	Machine to Human
ECG Sensor	Intelligent server	ECG is varying	Machine to machine
Intelligent server	Doctor	This diabetic patient will cause kidney disease	Machine to Human
Nutrition sensor	Patient	Nutrition is not sufficient	Machine to Human
Intelligent medicine box	Doctor	That medication is overdose	Machine to Human
Smart earrings	Patient/doctor	Your brain is not working	Machine to Human
Smart earrings	Patient/doctor	your ear is functioning Only 70%	Machine to Human
Smart ambulance	Hospital operation theatre	Patient present condition	Machine to machine
Doctor	Intelligent medicine box	Decrease the dosage of tablet	Human to machine

Fig. 1 Communication in H-IOT

doctor searches for the values of a body temperature (or any other vital value) in different situations of a particular patient. This is because he needs to diagnose a disease or to develop a new treatment mechanism. Is there any indexing technique available nowadays is suitable to index the correlated data value at the different times?

With the traditional indexing techniques used in search engines, viz., inverted index suffix trees, signatures files [1] are failed to establish a correlation among the documents. These techniques treat each and every document as a separate one. R tree, B + tree, and K-D tree are the well-known indexing techniques used to index spatial data. These all techniques fail to keep track of the records name of each data as well as the correlations among the records. According to [2], communication in H-IOT can be

a. Communication among IOT-MDs,
b. Communication between an IOT-MD and smart server, and
c. Communication between IOT-MD and human.

Message size cannot be determined in advance in each and every time. The messages can be information from the patient about his present symptoms or an instruction from the caregivers or it can be a discharge summary (Fig. 1).

2 Data Format in H-IOT

Named Data Networking (NDN) [3] is the future Internet for naming the data content in the Internet. Name of the data content is not visible to the network and is used for routing and forwarding the packets. Every content name may occupy the various lengths string arranged in a hierarchy. NDN contains three data structures. (1) Content store (CS)—The main purpose of CS is to store the previously accessed contents so as to satisfy to the future request. (2) Pending Interest Table (PIT)—If the content is not available in the CS, the content name is checked in the PIT, and it does not present, and then the requested content is added to its interface. The corresponding interest packet is forwarded to FIB. (3) Forwarding Information Base (FIB)-—Whenever FIB receives a content, new PIT entry is created. The longest prefix of the content name is calculated and the interest is forwarded to the outgoing interfaces of the FIB's longest prefix matching entry. Py-NDN API service is used for naming the healthcare IOT data.

IOT-MD as a single Vital Collector

Mid||Pid||TV||NV||VV||PLoc||Gid|Did||Dloc||TS||Message

Mid	Machine identifier
Pid	Patient identifier
TV	Type/name of vitals
NV	Number of values read
VV	Value of the vital.
Ploc	Patient location
Gid	Gateway ID
Did	Doctor ID
Dloc	Location of the doctor
TS	Timestamp of the vital read
Message	Any other message

Each value is separated by # symbol

IOT-MD as a Group of Vital Collector

Mid||Pid||TV|NVV||VV||PLoc||Gid|Did||Dloc||TS||Message

NVV Number of vital values read

To distinguish each vital, use the symbol $.

To distinguish the values within the vitals, use the symbol #.

IOT-MD as a Single Medication Supplier

Mid||Pid||Did||PLoc||TS||mid||Dn||Ds||Rm||Status||balance||

mid	Medication identifier
Dn	Duration of medication
Ds	Dosage of medication
Ds	Route of medication (e.g., IV)

IOT-MD as a group of Medication Supplier for a single doctor

Mid||Pid||Did||PLoc||TS||NM||mid_Dn_Ds_Rm||Status||balance

NM—Number of medications. Each medication is stored as a fourple format mid_Dn_Ds_Rm indicating the medication identifier, dosage, duration, and route of medication. Here, each record consists of data coming from an IOT-MD at an instant. Only vital value field is changing at the next instant.

3 Problem Definition

The massive amount of data produced by these Internet of Things-Medical Devices (IOT-MD) consumes a large amount of memory. Consider a scenario: We are monitoring the vital value of a heart diseased patient. Suppose the data coming from one IOT-MD at an instant need 128 bytes to store. The patient is monitored by an IOT-MD at every second, so it produces 128 * 60 bytes data at 1 min. If there are m number of IOT-MD monitoring a patient, every day, a single patient may produce 128 * 60 * 60/h * 24 * m bytes of data. If there are "n" of patients registered in a hospital, hospital needs 128 * 60 * 60/h * 24 * m * n bytes of storage every day. On the other hand, record coming from a particular IOT-MD, connected to the same patient at consecutive times, changes rarely. Even if there is a change, change will be in some fields, not on the entire record.

Hence, the problem is to build a suitable index structure for H-IOT, which is capable of updating some fields in the record frequently without updating the entire record while keeping all the previous field values for future reference.

4 Solution Approach

H-IOT documents consist of documents from IOT-MD as well as Electronic Health Records (EHR). The documents from an IOT-MD occupy both streaming data and the text messages. The following is the sample IOT-MD documents (Fig. 2).

M600343||P7890063||B.P||010||70#120||PL00234||G23412||D001234||D034562||12-6-2017:12:45:42:00PM||patient is anemic.

The text message can be any communication message from patient to a doctor or an instruction from doctor to a patient or communication from the data server to a patient

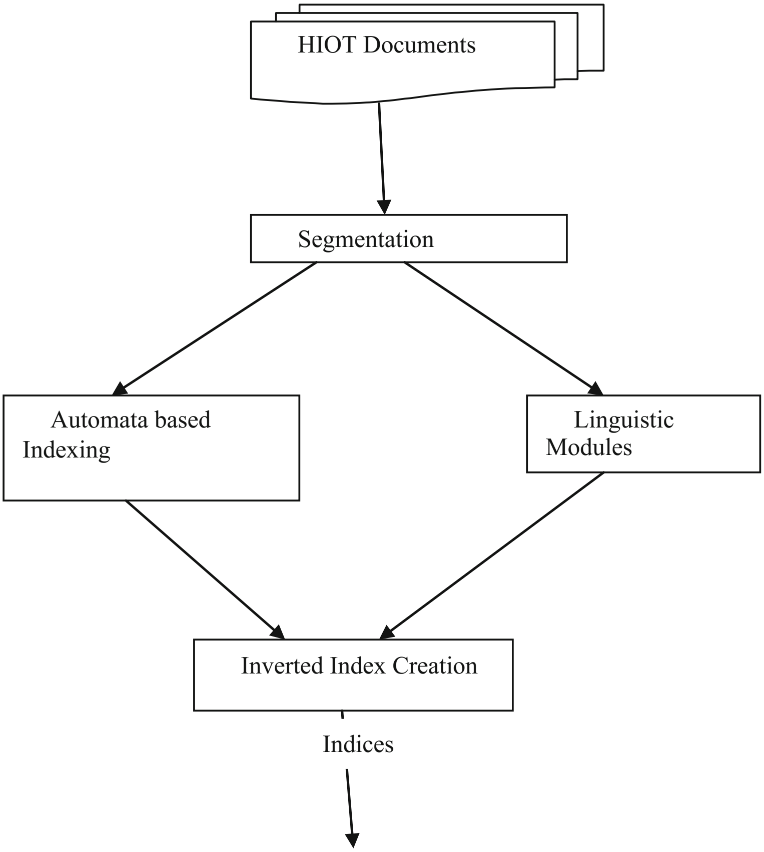


Fig. 2 Architecture diagram

or to a doctor. These text messages are undergone some linguistic expansion process. As these messages are medical text, medical ontology is required for expanding the medical terms, and the difference of term expansion strategy is represented in [4]. Streaming data from IOT-MD is given to an automata-based indexing module where automata indices are created. Results (concepts from the text messages) from the linguistic expansion module along with automata indices are used create inverted indices.

4.1 Automata-Based Indexing

Here, we create finite state automata [5] with the initial state as "Q0", and all other states are represented as triplet format containing the type of vital, number of vital,

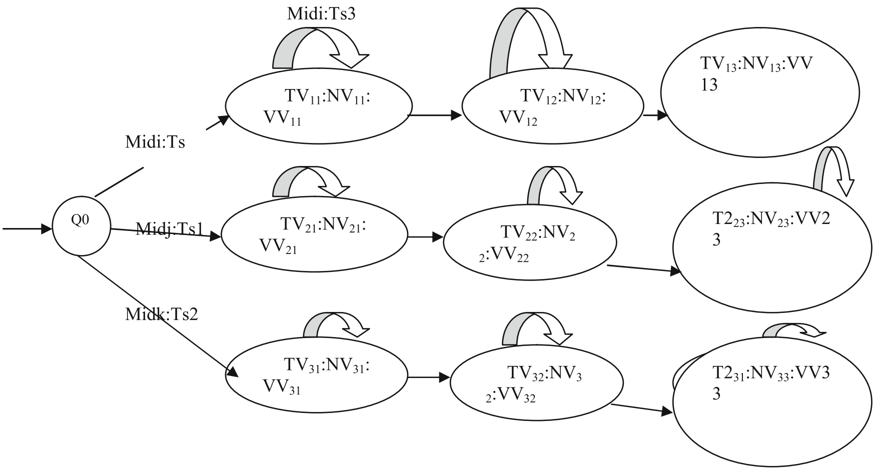


Fig. 3 Automata modeling

and value of the vital (TVi:NVi:VVi). Each value is separated by a # symbol. Each input is a twin pair containing machine identifier and timestamp (MIDi:TSi). The final state is marked as a double oval. Transition function maps from the triplet (TVi:NVi:VVi) X(SIDi:TSi) to (TVi:NVi:VVi) (Fig. 3).

Initially, we create an index table in which rows represent the triplet (TV:TSstart:TSEnd). TV is the name or type of the vital. TSstart is the timestamp of the TV value first recorded. TSEnd is the timestamp of the last TV value recorded. Columns contain a twin pair (pid:Rid), a patient identifier and a record identifier. The values in the matrix will be the vital value of a patient (pid) at the time interval (TSstart:TSEnd).

Once the index is created, all the incoming vital values are not needed to be stored, since many of them are same as the previous values. Only those vitals whose previous values are different from the previous one are needed to be updated. To find the relevant keyword, we construct an automaton from the previously created matrix.

Index Updation

```
Input: A Matrix , Record Ri,Vital value VVi,Timestamp
TSi
Output: Index Term
Extract patient identifier pid
For each patient identifier pid,
         For each vital type TV
         Find the vital value VV of the TV at the  last
         timestamp
If the vital values are not matching with the previous
value
         Create a new index term with (TV:TSi:TSi)
          Update the new value in the matrix with
         (TV:TSi:TSi) and (Pid:Ri) as VVi
Else  update the second timestamp of the previous index
term
         End for
End For
```

Searching
Searching for particular vital value of a patient is at an instant in O(1) time. If patient identifier is not given, search process takes O(n) time, where n is the number of distinct (Pid:Ri) pair. Searching for particular vital value of a patient at different times can be found in O(m), where m is the number of the distinct index terms (TV:TSi:TSi).

5 Results and Discussions

In our previous paper [4], we have mentioned some of the promising application areas of H-IOT. Based on these, we have created five different datasets, the first three datasets are the datasets for chronic patients and fourth and fifth are the datasets generated for well-being and elderly care. Datasets for chronic diseases are generated at specific intervals or when the request came from another side. Dataset for well-being is generated based on request upon patients or care providers or at particular period or when any abnormal events are detected by an IOT-MD.

In [4], they mentioned that the performance of the search system can be improved using concept-based retrieval. The various kinds of hypernymy, hyponymy, meronymy, and holonymy relations can be established and thus improvisations on the results can be accomplished.

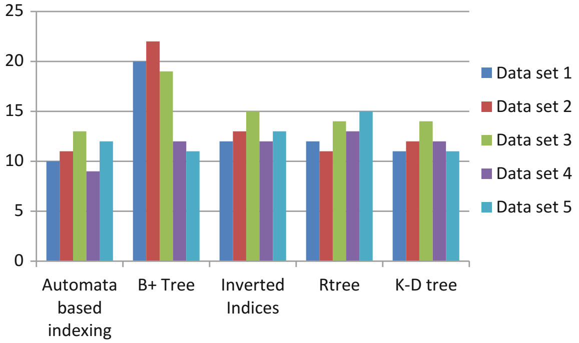


Fig. 4 Index storage space comparison chart

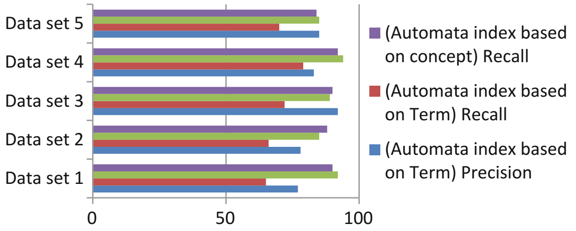


Fig. 5 Precision–recall chart

The generated Internet packet names are varying depending on the stakeholder it generated or it may depend upon the manufacturer. Suppose someone wants to retrieve the B.P value of a patient, he may be generated a packet like B.P//patientid//machine identifier OR blood Pressure//patientid//machine identifier OR CUI of blood pressure//patientid//machine identifier. UMLS ontology is used for resolving the ambiguous concept names by generating the concept unique identifier (CUI). Figure 4 represents the index storage space comparison of automata-based indexing with some existing indexing techniques. *X*-axis represents the different indexing techniques, and *y*-axis represents the storage space in megabytes (MB) (Fig. 5).

6 Conclusion

There has been a worldwide movement in e-medicinal services world utilizing the pervasive computing technology. Sequential scanning in very large databases may have a high response time. Unlike web resources, H-IOT has to deal with multiple kindle kinds of data. This has been necessitated by the invention of a new indexing mechanism in the H-IOT world in order to access the data rapidly. Storage space needed for the automata scheme is less when compared to the existing techniques. The precision–recall chart shows that concept-based automata indexing scheme yield better results.

References

1. Baeza-Yates, R., Neto, B.R.: Modern Information Retrieval, 3rd edn. Addison Wesley, Boston (1991)
2. Machine-to-Machine Communications (M2 M); Use Cases of M2 M applications for eHealth ETSI (2013)
3. Saxeena, D., Raychoudhury, V.: Radient: scalable, memory efficient name lookup algorithm for named data networking. J. Netw. Comput. Appl **63** (2016)
4. Sony, P., Suresh Kumar, N.: Concept based electronic health record retrieval system in healthcare IOT. In: International Conference on Cognitive and Soft Computing (2017)
5. Hopcroft, J.E., Motwani, R., Ullman, J.D.: Intro to Automata Theory, Languages and Computation, 2nd edn. Addison-Wisley, Boston (1979)

Monitoring Individual Medical Record by Using Wearable Body Sensors

Jagadeesh Gopal, E. Sathiyamoorthy and R. Mohamad Ayaaz

Abstract Every single essential sign has an immediate effect in human services, so body sensors are utilized for consistent social insurance checking. In existing techniques, the sensors are embedded or wear by the client and the sensor information are exchanged and put away in the restorative server. In this proposed framework is for a healing facility condition the specialists and attendants can access the client's therapeutic data from the Blynk server. The wearable sensor's on the patient's body is associated with an Arduino board the detected data is transmitted by transfer hub and put away in the cloud which will be access by the Blynk server and get the data to the Blynk application so that the patient and the doctors can access the data or the records of the particular patient and even the medical staff. The specialists and attendants can get to the patient's medicinal data from the Blynk application. Individual medical record [IMR] can be access by the doctors and the medical staff to analyze the patient condition.

1 Introduction

1.1 Background

In recent times, Individual Medical Records (IMR) has risen as a patient-driven model of wellbeing data trade. An IMR benefit permits a patient to make, oversee, and control it is close to home well-being information in one place through the web, which has made the capacity, recovery, and sharing of the restorative data more efficient [1]. A wearable body sensor network (WBSN) framework has been proposed to send in medicinal situations. The remote framework in the WBSN utilizes restorative groups to get physiological information from sensor hubs. The therapeutic groups are chosen to lessen the impedance and in this manner increment the conjunction of sensor hub gadgets with other system gadgets accessible at medicinal focuses. The gathered

J. Gopal · E. Sathiyamoorthy (✉) · R. Mohamad Ayaaz
School of Information Technology and Engineering, VIT, Vellore, India
e-mail: esathiyamoorthy@vit.ac.in

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_38

information is put away utilizing with the help IMR (Individual Medical Records) which is completely tolerant and controlled structure. The entryway hubs associate the sensor hubs to the neighborhood or the Internet [2]. In a doctor's facility zone to render nature of administration the framework comprises of a portable care gadget, which is in charge of catching and remotely sending the patient's ECG information, a wireless multi-hop relay network (WMHRN) that is accountable for handing off the information sent by the previous, and A residential gateway (RG), which is in charge of social event and transferring the got ECG information to the remote care server through the Internet to do the patient's well-being condition observing and the administration of neurotic information. A crisis ready administration utilizing short message benefit (SMS), in view of the location of irregular variety of HR, is additionally utilized as a part of the RG to further upgrade the medicinal services benefit quality. Hande Alemdar, Cem Ersoy in "Remote sensor systems for human services: A review" Says about the difficulties and different security issues in remote sensor systems, for example, security low power utilization, un-pretentiousness, adaptability, vitality effectiveness, security and give a thorough investigation of the advantages and difficulties of these frameworks [3].

1.2 Problem Statement

Each and every key sign specifically influences social protection, so body sensors are used for relentless restorative administrations watching. In existing procedures, the sensors are implanted or wear by the customer and the sensor data is traded through PDA (Personal Digital Assistant) and store in the restorative server. In proposed technique, Individual Medical Records (IMR) comes into play. Individual Medical Records (IMR) is totally tolerant controlled system which gives extra security and flexibility. This information is secured on outcast servers, i.e., cloud advantage providers. Nowadays there are a lot of security stresses over all IMRs from unapproved get to. Remembering the true objective to get to our own particular IMRs, it ought to be mixed before securing in cloud. Still there are a couple of troubles, for instance, trustworthiness, security, adaptability which needs to overcome by any capable structure. In this paper, we propose a novel patient-driven framework for data get the opportunity to control to IMRs set away in semi-confided in Blynk servers. To achieve our goal of most secure and versatile structure we are accepting credit-based encryption to encode patient's record. In our structure, we have concentrated on multi-characteristic records from different substances having a place with different spaces. Patient's data security is kept up by using multi-quality records in secure regions. This will similarly give on demand customer credit revocation to emergency staff in veritable conditions. Moreover, we are endeavoring to make the structure affirmation in disconnected condition.

2 Proposed Work

In the proposed system, body sensors been connected to the patient's body. Using Arduino, data can be collected from the patient's body. A Wi-Fi module is used to transmit the data to the application from which the doctors and the patients can access the particular data. For the framework segments to work, we are utilizing a microcontroller called Arduino to control the wearable body sensor and the Wi-Fi module called ESP8266. The Wi-Fi module is utilized to associate the framework to the Internet and send the notification and the data to the client. At the point when the conditions of the sensor sense the data, it will flag the microcontroller which will thus instate the Wi-Fi module. We have utilized the Blynk application as an extension between the Wi-Fi module and the Internet. The API given by the Blynk application sends the information gathered from the client to the client. The Wi-Fi module likewise can acknowledge charges from the Blynk application which can be utilized to send the data on or remotely by utilizing the Blynk cell phone application.

3 Materials and Methodology

Hardware apparatus required for this venture of idea are Arduino UNO, Temperature sensor, Smoke sensor, Wi-Fi module [ESP8266], a breadboard and few jumping wires. Likewise software required to get an output from the PC, Arduino IDE is an open source software used to upload the standard code of the sensors and the Arduino connected to the PC with the help of USB cable. Apart from the Arduino IDE, Blynk application is used to get the data from the sensors that are sensed from the human body. Blynk is a Platform with iOS and Android applications to control Arduino, Raspberry Pi and the inclinations over the Internet. It is a propelled dashboard where you can amass a sensible interface for your wander by basically moving devices. It is really simple to set everything up and you will start tinkering in less than 5 min. Blynk will set you up on the web and for the Internet of Things.

To implement this IoT concept, myself using Blynk libraries that takes information from the microcontroller and sends the information to the host (Blynk application) by utilizing the implicit capacities and methodology of the Blynk libraries. The application acts as a virtual controller for our Arduino microcontroller which can get information from the framework and send summons to it utilizing the Wi-Fi module as a channel. The application creates a validation id for each new venture which is utilized by the microcontroller to speak with the Blynk application over the web.

It is all around perceived that the IoT innovation has been picking up energy and is being coordinated in ordinary things, for example, cars, electrical machines, and so forth. A zone where IoT is being considered as one of the urgent advancements is for security purposes. Healthcare system is an important area of concern for all the doctors and the patients whose data can be retrieving through the concept of IoT. This concept is made for security reasons to store the patient record in an effective manner so that the third party cannot access the data without the permission of the particular patient.

4 System Architecture

Architecture diagram shows that the Arduino UNO leading the process by allowing the sensors to connect in it. Arduino UNO been connected to PC by using the USB cable. ESP8266 (Wi-Fi module) also connected to the Arduino, it will transmit the sensed data to the Blynk server from which user will get in the Blynk application. Engineering is connected to comprehend the usefulness of the venture. Architecture clarifies the availability procedure between all sensors to microcontroller to Wi-Fi module to Blynk server to Blynk application. Architecture likewise shows the stream of correspondence between the parts and the different qualities in an effectively justifiable organization (Fig. 1).

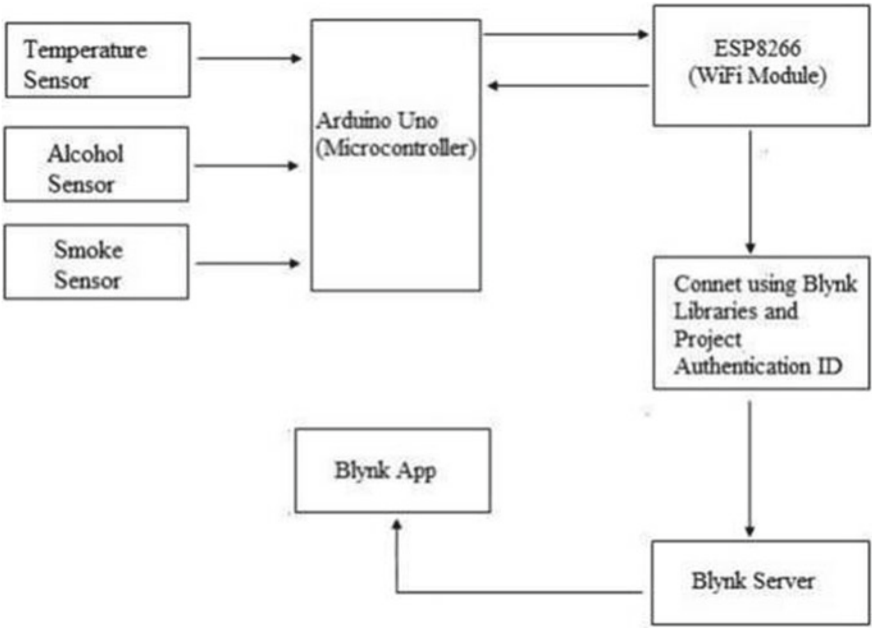

Temperature Sensor
Alcohol Sensor
Smoke Sensor
Arduino Uno (Microcontroller)
ESP8266 (WiFi Module)
Connet using Blynk Libraries and Project Authentication ID
Blynk App
Blynk Server

Fig. 1 Architecture diagram

5 System Implementation

5.1 *Setting up ESP8266*

Before we start associating the equipment, we need to get the ESP8266 set up by blazing the most recent rendition of the firmware accessible for the module. This is on the grounds that the chip accompanies a more seasoned form of the AT summon firmware pre-introduced out of the crate which cannot speak with the Blynk libraries productively and will give a mistake with our code. In this way before we start, download the ESP8266 flasher instrument and the most recent firmware from the web which would be in the receptacle arrangement and set up the ESP8266 to the Arduino Uno (Fig. 2).

5.2 *Hardware Configuration*

Once the ESP8266 has been flashed with the most recent firmware, we are prepared to amass our venture parts together. For this we will require a breadboard to associate the microcontroller, reed sensor, signal and the ESP8266 utilizing the jumper wires. The breadboard is utilized to interface between the different segments accessible. It additionally makes it simple to associate different contributions to a solitary stick on the Arduino board. The stick setup for all parts and configuration draw taking after the real circuits.

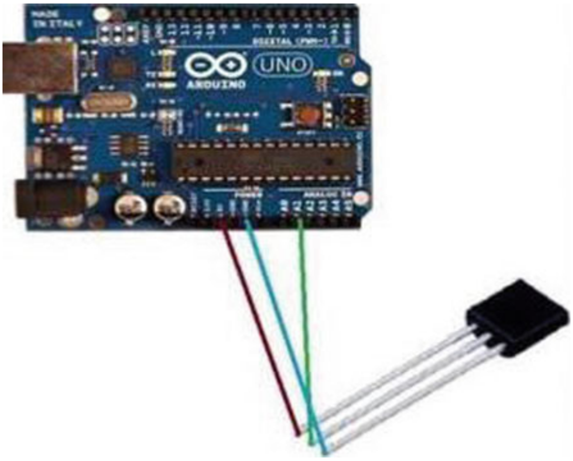


Fig. 2 Circuit diagram (design check)

5.3 Circuit Diagram

Taking after portray which has been developed utilizing the Fritzing programming demonstrates how the segments should be associated together utilizing the breadboard and the jumper wires. The last arrangement requires not be indistinguishable to the given outline, despite the fact that the pins on every gadget should be associated with the same comparing pins on the Arduino Uno board.

5.4 Configure the Blynk Application

Creating an account

- After installing the Blynk app on either android or the iOS devices, you need to create an account to access the services. Initially it will ask for Log-in or Create new account. Opt for Create new account as a new user.
- After the entering the email and password, click on the Sign Up button. It will create the account and user will sign in automatically.

Creating Project

- Once the user have logged in, user can click on the Create new project. This will take the user to Project design screen wherein the client needs to pick the title of the venture, the sort of gadget which the application will speak with and the method of the correspondence.
- For our venture, we select Arduino Uno as our gadget and Wi-Fi as the association sort. Client can likewise choose the choice to pick among dim and light subject according to their inclination which will just change the presence of the application and not influence with any venture alternatives.

Authentication Code

- After you have made the venture, the Blynk App will create an Authentication Token for your venture. Keep in mind that each venture has an alternate Authentication Token so that Blynk servers can recognize them thus that your code transferred to your gadget knows which venture to convey to. The confirmation token is sent to the email-id utilized while joining and can likewise be seen inside the venture settings screen.

Adding Widgets to the Project

- To interface with our segments, we have to add Widgets to our venture. To include Widgets tap the "+" catch, this will draw out the gadgets sheet from the correct corner of the screen which contains the rundown of all gadgets accessible. Simply tap on the gadget you have to add it to your venture.
- According to our prerequisites, we include a Notification gadget and a Button gadget to our venture. The notice gadget is activated when the tell() technique is called from the program and is utilized to show Push warnings on iOS gadgets. The catch is added to control the conduct of the ringer and will be utilized to stop or begin the signal remotely from inside the application.

6 Result/Output

6.1 Upload the Code

The code need to be compiled once it done, upload it to the Arduino UNO board by connecting the board to the computer using USB cable.

6.2 Running the Program

Subsequent to transferring the program, tap on "Serial Monitor" to begin running the code. Once the code begins to run, the principal thing it will do is to attempt and interface the ESP8266 to the Access Point pre-characterized in the code. On the o chance that the ESP8266 interfaces with the Access Point with web abilities, it will associate with the venture by means of the Blynk Servers by sending a ping message.

6.3 Getting Data/Output

The user will get the temperature in the Blynk application which is retrieved from the temperature sensor via Arduino.

7 Conclusion and Future Scope

7.1 Conclusion

The hardware and software both works perfectly as they were expected to with no errors. From sensing it from the patient body to the retrieving the circumstances and sending it to the Blynk application so that the doctors and medical stuff can view it. Even the patient also can view it using their smartphone. Another imperative part of the venture is the availability between the ESP8266 (Wi-Fi module) and the Blynk server. The framework effectively associated with the Blynk server utilizing the confirmation token and the Blynk libraries. Subsequently, we could get the warning on our cell phones when there was any adjustment in the status of the reed module sensor. Likewise the extra capacity to control the alert remotely is exceptionally useful and can be extremely valuable in some unexpected conditions. It was additionally watched that the Blynk application worked easily and done all correspondence between the equipment and the application precisely. In the wake of testing the framework ceaselessly at an interim of 2 h for a time of 10 h, amid which the equipment was not turned off even once, yielded similar outcomes which was that every one of the functionalities were working splendidly in sync. This was done to test the perseverance of the equipment and to test the reliance of the product. The test was furnished with every one of the suppositions being valid for the entire length of time which implies that the cell phone and get to point don't lose the web association and the power supply to the Arduino Uno load up is not cut off.

7.2 Future Scope

The medical Internet of things will play a major role in the medical world. In future we improvise our venture with the help of ready message can be send to specialist when a parameter achieves an unusual esteem. And this proposed framework can be utilized by the healing facility of clinical range to enormous doctor's facility condition to utilize the cloud administrations like as web administrations.

References

1. Ibraimi, L., Asim, M., Petkovic, M.: Secure management of personal health records by applying attribute-based encryption. In: Wearable Micro and Nano Technologies for Personalized Health (pHealth), pp. 71–74 (2009)
2. Li, M., Yu, S., Cao, N., Lou, W.: Authorized private keyword search over encrypted data in cloud computing. In: International Conference on Distributed Computing Systems, pp. 383–392 (2011)

3. Lai, C.-C., Lee, R.-G., Hsiao, C.-C., Liu, H.-S., Chen, C.-C.: A H-QoS-demand personalized home physiological monitoring system over a wireless multi-hop relay network for mobile home healthcare applications. J. Netw. Comput. Appl. **32**, 1229–1241 (2009)
4. Li, M., Yu, S., Ren, K., Lou, W.: Securing personal health records in cloud computing: patient-centric and fine-grained data access control in multi-owner settings. SecureComm. **10**, 89–106 (2010)
5. Lohr, H., Sadeghi, A.-R., Winandy, M.: Securing the e-health cloud. In: International Health Informatics Symposium, pp. 220–229 (2010)
6. Mandl, K.D., Markwell, D., MacDonald, R., Szolovits, P., Kohane, I.S.: Public standards and patients' control: how to keep electronic medical records accessible but privateMedical information: access and privacy Doctrines for developing electronic medical records desirable characteristics of electronic medical records challenges and limitations for electronic medical record-s conclusions commentary: open approaches to electronic patient records commentary: a patient's viewpoint. Bmj **322**, 283–287 (2001)
7. Benaloh, J., Chase, M., Horvitz, E., Lauter, K.: Patient controlled encryption: ensuring privacy of electronic medical records. In: ACM Work-shop on Cloud Computing Security, pp. 103–114 (2009)
8. Yu, S., Wang, C., Ren, K., Lou, W.: Achieving secure, scalable, and ne-grained data access control in cloud computing. Infocom, pp. 1–9 (2010)
9. Dong, C., Russello, G., Dulay, N.: Shared and searchable encrypted data for untrusted servers. J. Comput. Secur. **19**, 367–397 (2011)
10. Goyal, V., Pandey, O., Sahai, A., Waters, B.: Attribute-based encryption for ne-grained access control of encrypted data. In: ACM Conference on Computer and Communications Security, pp. 89–98 (2006)
11. Li, M., Lou, W., Ren, K.: Data security and privacy in wireless body area networks. IEEE Wireless Commun. **17** (2010)
12. Boldyreva, A., Goyal, V., Kumar, V.: Identity-based en-cryption with e cient revocation. In: ACM Conference on Computer and Communications Security, pp. 417–426 (2008)
13. Ibraimi, L., Petkovic, M., Nikova, S., Hartel, P., Jonker, W.: Ciphertext-Policy Attribute-Based Threshold Decryption with Flexible Delegation and Revocation of User Attributes. Technical Report, Centre for Telematics and Information Technology, University of Twente (2009)
14. Yu, S., Wang, C., Ren, K., Lou, W.: Attribute based data sharing with attribute revocation. In: ACM Symposium on Information, Computer and Communications Security, pp. 261–270 (2010)
15. Narayan, S., Gagne, M., Safavi-Naini, R.: Privacy preserving EHR system using attribute-based infrastructure. In: ACM Workshop on Cloud Computing Security Workshop, pp. 47–52 (2010)
16. Liang, X.-H., Lu, R.-X., Lin, X.-D., Shen, X.S.: Patient self-controllable access policy on phi in ehealthcare systems. AHIC, pp. 1–5 (2010)
17. Bethencourt, J., Sahai, A., Waters, B.: Ciphertext-policy attribute-based encryption. Security and Privacy, pp. 321–334 (2007)
18. Akinyele, J.A., Pagano, M.W., Green, M.D., Lehmann, C.U., Peterson, Z.N.J., Rubin, A.D.: Securing electronic medical records using attribute-based encryption on mobile devices. In: ACM Workshop on Security and Privacy in Smartphones and Mobile Devices, pp. 75–86 (2011)
19. Darwish, A., Hassanien, A.E.: Wearable and implantable wireless sensor network solutions for healthcare monitoring. Sensors **11**, 5561–5595 (2011)
20. Kumar, R., ManjupPriya, S.: Cloud based M-Healthcare emergency using SPOC. In: Advanced Computing (ICoAC), pp. 286–292 (2013)

CLOUD COMPUTING-TMACS: A Robust and Verifiable Threshold Multi-authority Access Control System in Public Cloud Storage

K. Selvakumar, L. SaiRamesh, S. Sabena and G. Kannayaram

Abstract Cloud Computing is a process of storing and processing data. However, that data can be vulnerable to be hacking and data manipulation, this paper focuses on the security and privacy of the cloud and what methods we use to do it which involves data encryption and decryption. This paper focuses on a public key encryption algorithm called ABE (Attribute-Based Encryption) which have classified as Key-Policy based ABE (KP-ABE), Cipher text based ABE (CP-ABE), and Multiple authority Attribute-based encryption (MA-ABE). In this case, we choose the CP-ABE public access control scheme for storing data in the cloud called TMACS. Here this scheme takes the advantage of (t, n) threshold secret sharing i.e. number of authorities' share a master key and jointly deal with a costume attribute become hard inside which each ability is gifted in the direction of release attributes independently. The user must contact any t $(t < n)$ authorities to produce their respective private key. This solves the problem of having only one authority to have the attribute set as it may cause single-point hindrance to performance and security.

1 Introduction

Cloud computing is the act of utilizing a system of remote servers facilitated on the Internet to store, oversee, and process information. The cloud is shabby to utilize, secure, adapts to request and aides in decreasing carbon impression. Though advantageous the cloud computing can be the security and privacy of owner's data is major concern in cloud storage [1, 2]. Therefore, secure data access is critical

K. Selvakumar (✉) · G. Kannayaram
School of Computer Science and Engineering, VIT University, Vellore 632014, Tamil Nadu, India
e-mail: kselvakumar@vit.ac.in

L. SaiRamesh
Department of Information Science and Technology, College of Engineering Guindy,
Anna University, Chennai 600025, India

S. Sabena
Department of CSE, Anna University - Regional Centre, Tirunelveli 627007, India

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_39

issue in cloud storage. Various techniques to secure data in the cloud exist, but the ones we take on in this paper is a public key encryption algorithm: Attribute-Based Encryption (ABE). We do not take symmetric key encryption into consideration as the key which encrypts and the message is sent together from sender to receiver, this way vulnerability exists. In public key encryption, the sender will send the encrypted message with their public key, the receiver will have to decrypt the message with their respective, personal private key.

ABE encryption technique is not the only technique in PBE, there is IDE (Identity-Based Encryption) and Fuzzy-IDE as well. In every PBE, there is always a PKG (Private Key Generator) which uses the user's identity (such as e-mail) as a public key to encrypt the message if the user wishes to send and produces a private key for decryption of received messages [2, 3]. IDE follows the said algorithm and it is the simplest PBE algorithm, however to make it more secure the Fuzzy-IDE introduces a tolerance factor d (integer) which restricts public users to send data, here another step is introduced each user must have a set of attributes, this set constitutes the identity of a user, so whenever two users would like to send data to each other, the number of common attributes among the two must be greater than or equal to the tolerance factor of the receiver. Now finally we reach the (ABE) algorithm, here, the decryption of an encrypted message or Cipher text is conceivable just if the arrangement of traits of the client key matches the properties of the figure content [3, 4]. Unscrambling is just conceivable when the quantity of coordinating qualities is no less than a limit esteem.

ABE algorithms have been chosen for this paper as they have fruitful advantages such as: data confidentiality, secured access control, scalability, accountability, revocation and collision resistant. Variations of the ABE algorithms have produced multiple schemes such as KP-ABE, CP-ABE and MA-ABE [5, 6]. Here amongst the said schemes, CP-ABE is picked among the rest as the entrance arrangement is related with figure content and the quality set are related with private key. Also, it works well with the (t, n) threshold secret sharing scheme.

However, they are some limitations: the efficiency, secured access control, computational overhead of CP-ABE is normal, unscrambling keys just help client qualities that are sorted out consistently as a solitary set, so the client can just utilize every single conceivable mix of properties in a solitary set issued in their keys to fulfill arrangements, along these lines demonstrating that this plan cannot hold to client disavowal and responsibility.

2 Background Knowledge

Now, to explain the (t, n) threshold secret sharing scheme. Suppose a dealer decides to share a secret to n number of people. In order to keep the secret safe, the secret is split amongst n men (E.g.: giving four random digits in a 20-digit combination lock.). Now it is not necessary that all n men have to gather to unlock the secret, only a certain number is required to know the secret. Any number below that will not be sufficient to know the secret, this number is called threshold.

Now how this plays with CP-ABE is that the CA (Certified Authority) will generate a master key and provide it to all the attribute authorities of that system. So, when a user tries to generate his/her secret/private key, it has to be produced with more than *t* authorities. This framework TMACS (Threshold Multi-Authority Access Control System) will be undeniable secure and powerful when not as much as ***t*** experts are available in the framework.

3 Literature Survey

Dispersed, Concurrent and Independent access to encoded cloud databases gives numerous, free and disseminated customers to execute simultaneous inquiries on scrambled information. Here even SQL explanations are changed into scrambled structures to give privacy [6]. Here SecureDBaaS process plaintext information, encoded information, metadata and scrambled metadata. SecureDBaaS users can obtain the metadata from cloud through SQL statements. The problem with this approach is that all the SQL commands types need to be predefined during design phase which seems impractical.

Adaptive encryption architecture for cloud gives the owner privileges to change the set of SQL queries even after database design. Encrypted engine requires metadata to execute SQL queries from the cloud database and decrypt it through private key which is with client side application. Adaptive encryption scheme consider many SQL aware encryption algorithm such as Random & Deterministic, sums and search.

Here, encryption calculations are sorted out into structures called onions, where every onion is comprised of a requested arrangement of encryption calculations called layers. Each plaintext column is corresponding to an onion. This architecture called Multi user relational encrypted database that assures data privacy by executing SQL operation [7, 8]. Here data managed and create by database authority, who is also responsible storing encrypted data and metadata on the cloud. Each user will be given some credentials including the information that allows him/her to access legitimate data. The database authority is the only person who can control all system entities; this can leads to the database authority overloading and can result on performance degradation.

The step-by-step process to get to control is a critical issue in testing in the distributed framework while getting to control in a private cloud is easier. A lot of work has been put into improve these ABE plans and eliminate every bottleneck that has risen. In these multi-expert access control plots, the entire characteristic set is partitioned into numerous disjoint subsets and kept up by various specialists, yet each quality subset is separately kept up by one and just a single expert, which influences the issue of single point to bottleneck on security still exist in frameworks [9, 10]. Since having just once Cloud Authority (CA) which could turn out to be a bottleneck, Chase has proposed the vital multi-specialist ABE plot, in which an overall affirmation control (CA) is introduced. There are also ABE plans which do not involve the CA for example KP-ABE. In this paper, we propose a multi-control CP-ABE (Bit Exchange get to control design) design, which deals with the single-point bottleneck on security.

In this arrangement, no one has full control of a trademark, there are various forces who have control over the subset. In CP-ABE outlines, there is a dark key used to make private keys like itself. The likelihood of (t, n) sharing guarantees that the dark key cannot be gotten by a specific power alone. This course of action is not just evidently secure when not as much as t powers are considered, in like manner when not as much as t powers are in the framework. This course of action is the basic undertaking to address the single-point bottleneck on security in CP-ABE outlines in scattered disseminated stockpiling.

4 Existing System

Existing System boasts an Attribute-based Encryption (ABE) which is a champion among the most proper plans to control data access with no attempt at being subtle fogs since it can ensure data proprietors arrange control of their data and give a point by point get the opportunity to control profit. An Attribute-based Encryption (ABE) is separated into two classifications, for example, Key-Policy Attribute-based Encryption (KP-ABE) and Cipher content Policy Attribute-based Encryption (CP-ABE). Conversely with KP-ABE, CP-ABE is a favored choice for laying out access control for open dispersed stockpiling. In existing CP-ABE plots only a solitary master accountable for property organization and key appointment. This unrivaled expert circumstance can bring a single-point bottleneck on security. Yet some multi-master CP-ABE designs have been proposed, in any case they can't deal with the issue of single-point bottleneck on both security and execution.

4.1 Disadvantages of Existing System

In CP-ABE plan, one and only a solitary master is accountable for managing the qualities and key transport. Once the expert is exchanged off, any individual can without quite a bit of an extend get some individual's energy's master key, and thereafter he/she can create private keys of any attribute subset to unscramble the encoded data record. Along these lines, this unrivaled expert circumstance can bring a lone point bottleneck on both security and execution.

In multi-specialist CP-ABE conspire, the individual can get private keys of particular properties by bargaining at least one experts. In this way, the single-point bottleneck on execution and security is not yet comprehended.

4.2 Proposed System

In this paper, we propose an intense and evident utmost multi-specialist CP-ABE gets the opportunity to control intend to deal with the single-point bottleneck on both security and execution existing plans. In this procedure, different specialists commonly manage the whole quality set yet no one has full control over a specific property. In this way, we use TMACS (Threshold Multi-Authority Access Control System) nearby CP-ABE.

5 System Architecture

5.1 A Global Certificate Authority (CA)

This worldwide endowed substance in the framework is for the most part responsible for verification and additionally account actuation. It sets up framework parameters and in addition characteristic Public Key (PK) of each trait of the entire property set and subsequently is in charge of the development of the framework. By allotting an interesting guide to each legitimate client and every AA, CA acknowledges clients and Attribute Authorities (AAs') enrollment asks for as appeared in Fig. 1.

5.2 Multiple Attribute Authorities (AAs)

The main functions for the attribute authorities are authentication, key generation and forwarding a private key to the data user based on the user request. AAs can be overseers or directors of the application framework and consequently partake in the obligation to build the framework. Each AA generates a corresponding secret key independently, when it comes to generating the user's secret key.

5.3 Data Owners

The data owner contains all the data that is stored in the cloud and that is meant to be shared. The data owner has various functionalities. It is responsible for authentication and data encryption of a file and an access policy is defined by the data owner about who can get access to his/her data. Symmetric encryption algorithms such as AES or DES are implemented to encrypt his/her data. It also sends a key request to the CA and then receives the public key from it.

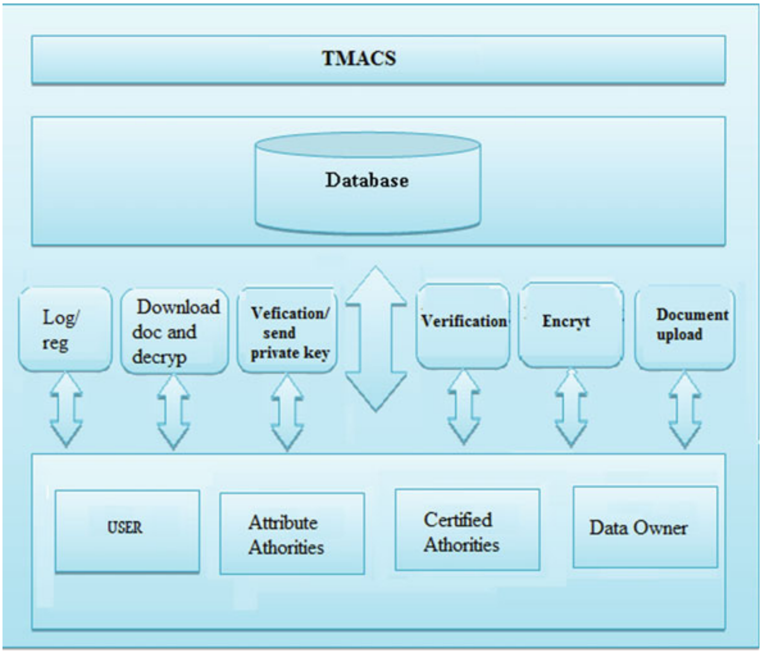


Fig. 1 Robust and verifiable threshold multi-authority CP-ABE access control scheme

5.4 *Data Consumers (Users)*

The data user is a client for the data owner. Permission must be taken from the attribute authorities by the user in order to access the data stored in the cloud. The data user is in charge of authentication, sending a key request, collecting the private key and decrypting data. To access the detail that is present in the system, the users have authentication and security for the following module.

5.5 *The Cloud Server (Database)*

The cloud server is a vast platform where all the data is stored by the data owner and this encrypted data can be shared as well hence this becomes a vital administration of distributed computing and it additionally gives benefits the information to be outsourced to store in cloud by means of web by the information proprietor. The cloud supplier deals with the cloud server and the cloud server is constantly on the web. The information customers are allowed to download the encoded information from the cloud.

6 Implementation and Result Analysis

The experiments are carried using CloudSim simulator where the 10 users are initialized and maintains the separate server in different virtual machine within the same host machine. The certificate authority is maintained at another virtual machine to monitor or maintain the proper access control in overall cloud environment. Performance is measured based time taken for encryption and decryption with certificate verification and detection accuracy of unauthorized access. The proposed TMACS is compared with [11, 12] to show the better performance.

Tables 1 and 2 show the time taken for encryption and decryption by different system is compared with proposed TMACS. The output show that our proposed system is performs 33% better than existing system. Even though the file size may differ but the performance of our proposed TMACS does not get degraded.

Apart from the time comparison, main thing is how the access controls is provided by TMACS and maintain the secure transactions.

Table 3 shows the experiment results with different number of unauthorized user's access and verify the number unauthorized users detected by the proposed system in different time intervals. First experiment is executed in five minutes time interval, second experiment in 15 min, third experiment in 30 min, fourth experiment to have 30 unauthorized access within 45 min and finally fifth experiment in one hour. The detection accuracy of TMACS is better than existing system described in [11, 12].

Table 1 Time taken for encryption of different files

File size (KB)	Time taken (ms)		
	[11]	[12]	Proposed TMACS
32	18	16	17
64	22	21	19
128	28	25	20
256	38	28	20.5
512	45	32	22

Table 2 Time taken for encryption of different files

File size (KB)	Time taken (ms)		
	[11]	[12]	Proposed TMACS
32	15	13	13
64	17	15	13
128	21.5	18	14.5
256	30	23.5	16
512	35	26	17.75

Table 3 Performance of TMACS on access control

No. of unauthorized access	Detection accuracy (%)		
	[11]	[12]	Proposed TMACS
5	98	98	98
10	95	96	97
20	88	92	95.5
30	81	89	94.7
50	76	85	93

7 Implementation and Result Analysis

In this proposed model, a new access control scheme called TMACS for providing secure data access in public cloud storage. Multiple attribute authorities generate the master key and share the key among the multiple users. A secret key is generated to authorize the permissible user to access the shared data by applying the threshold (t, n).

The result analysis shows the access control scheme proposed in this system is better than other encryption techniques. We can undoubtedly discover suitable estimations of (t, n) to influence TMACS to secure when not as much as t masters are exchanged off and solid when no not as much as t specialists are alive in the structure. Also, TMACS avoid the bottleneck problem by shared the data among multiple users.

This scheme implements a promising technique and subsequently it can be connected in any remote stockpiling frameworks and online informal organizations and so on. In future, optimized interaction protocols will be used for reducing delay in data sharing and we may incorporate Certificate Authority for key generation and verification instead only for certificate issuing.

References

1. Hur, J.: Improving security and efficiency in attribute-based data sharing. IEEE. Trans. Knowl. Data Eng. **25**(10), 2271–2282 (2013)
2. Kui, R., Cong, W., Qian, W.: Security challenges for the public cloud. IEEE. Internet Comput. **16**(1), 69–73 (2012)
3. Liang, K., Fang, L., Susilo, W., Wong, D.: A cipher text-policy attribute-based proxy re-encryption with chosen-cipher text security. In: 5th IEEE International Conference on Intelligent Networking and Collaborative Systems (INCoS'13), pp. 552–559 (2013)
4. Wang, Y., Chen, K., Long, Y., Liu, Z.: Accountable authority key policy attribute-based encryption. Sci. China Inform. Sci. 55(7), 163–1638 (2012)
5. Yang, K., Jia, X.: Third-party storage auditing service. Security for Cloud Storage Systems, pp. 7–37. Springer, New York (2014)

6. Zu, L., Liu, Z., Li, J.: New cipher text-policy attribute based encryption with efficient revocation. In: IEEE International Conference on Computer and Information Technology (CIT'14), pp. 281–287 (2014)
7. Wan, Z., Liu, J., Deng, R.: Hasbe: a hierarchical attribute-based solution for flexible and scalable access control in cloud computing. IEEE Trans. Inf. Forensics Secur. **7**(2), 743–754 (2012)
8. Yang, K., Jia, X.: Expressive, efficient and revocable data access control for multi-authority cloud storage. IEEE Trans. Parallel Distrib. Syst. **25**(7), 1735–1744 (2013)
9. Wu, Y., Wei, Z., Deng, H.: Attribute-based access to scalable media in cloud-assisted content sharing. IEEE Trans. Multimedia **15**(4), 778–788 (2013)
10. Lewko, A., Okamoto, T., Sahai, A., Takashima, K., Waters, B.: Fully secure functional encryption: attribute-based encryption and (hierarchical) inner product encryption. In: Gilbert, H. (ed.) Advances in Cryptology–EUROCRYPT 2010. LNCS, vol. 6110, pp. 62–91. Springer, Berlin, Heidelberg (2010)
11. Jayakumar, J., Sai Ramesh, L., Pandiyaraju, V., Muthurajkumar, S., Rakesh, R.: Secure data storage in cloud using decentralized access control in cloud. Proc. Adv. Nat. Appl. Sci. **9**(6), 192–196 (2015)
12. SaiRamesh, L., Sabena, S., Thangaramya, K., Kulothungan, K.: Trusted multi-owner data sharing among dynamic users in public cloud. Aust. J. Basic Appl. Sci. **10**(2), 315–319 (2016)

Analysis of CPU Scheduling Algorithms for Cloud Computing

Ramasubbareddy Somula, Sravani Nalluri, M. K. NallaKaruppan, S. Ashok and G. Kannayaram

Abstract Cloud computing refers to the advancement of distributed computing which takes the computational aspects of data processing into high-power centralized data centers over networks. It refers to the use of a centralized pool of resources which are allocated to a large number of customers on a pay-as-you-go model. This creates the need of scheduling algorithms which allow us to decide which job process, amongst the ones received will be allocated resources first for execution. Job scheduling is one of the major areas of research nowadays in the field of cloud computing as it helps not only in proper utilization of resources but also in avoidance of deadlock within the cloud infrastructure. There exist a large number of scheduling algorithms such as First Come First Serve Scheduling Algorithm, Generalized Priority Algorithm, Round Robin Scheduling Algorithm and Least Slack Time Scheduling Algorithm. This paper proposes a mechanism to implement each of these and then compute their average waiting times and average turnaround times for a common test case is more efficient one for cloud environment.

1 Introduction

Scheduling is one of the main and most important functions of an operating system. It allows for the fair and efficient usage of computing resources by all the waiting processes in a computing environment. In the cloud computing narrative, the usage of scheduling algorithms becomes all the more important due to the presence of a large number of computing resources which are all integrated into a single computing network to be able to behave and operate as a single computing resource of higher capabilities and capacity. Whether standalone computers or supercomputers connected via the use of a network, the need of deciding which process to run first always is the most fundamental procedure which needs to be carried out. As the number of computing resources available increase, so do the number of processes

R. Somula (✉) · S. Nalluri · M. K. NallaKaruppan · S. Ashok · G. Kannayaram
VIT University, Vellore 632 014, Tamil Nadu, India
e-mail: svramasubbareddy1219@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_40

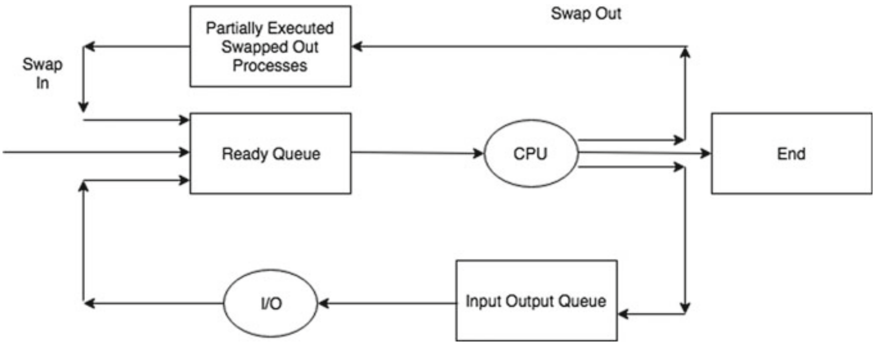


Fig. 1 Scheduling process

available for execution. As the number of processes available increases by leaps and bounds, one of them must be chosen for execution. The process of choosing one of these processes for execution is referred to as scheduling and the algorithm used for this selection is referred to as the scheduling algorithm. The scheduling process described in Fig. 1.

All Operating Systems Have the Following Three Types of Schedulers

- Long-term scheduler: The long-term scheduler is basically used to control the degree of multiprogramming. It is also responsible for deciding the degree of concurrency to be supported by the CPU [1], which allows for a large number of processes to be executed by the system. In operating systems nowadays, it is incredibly important for the efficient usage of the long-term scheduler as it allows for real-time processes to be executed in a time-bound manner. It also helps in deciding how to split computing resources between input output intensive and CPU-intensive processes. It decides whether a process is allowed to be admitted into the ready queue.
- Short-term scheduler: The short-term scheduler is used mainly to pick one of the processes from the ready queue and assign it some computing resources for execution. It works faster and also more than the long term scheduler as well as the medium-term scheduler. It is also capable of preemptive scheduling which allows it to 'force-out' processes from the CPU for other processes which it deems to be more important [1].
- Medium-term scheduler: The mid-term scheduler plays the role of a preemptive scheduler which allows it to swap out processes from the CPU when they either take too much time or memory in favor of processes which have higher priority. They then swap the process back into the CPU after execution of higher priority processes.

The following swapping of processes are performed by the schedulers mentioned above

1. Running State to Waiting State
2. Running State to Input Output Waiting State
3. Ready State to Running State
4. Process Termination.

2 Background

2.1 Scheduling Algorithm

I. **First Come First Serve**

This type of scheduling algorithm is the simplest of all the algorithms that exist [2]. Essentially, it is an implementation of a standard queue data structure, where-in processes are enqueued from the back of the queue and are dequeued as and when the need arises from the front. The processes are enqueued as and when they arrive in the ready state. The processes are dequeued and enter the running state sequentially in the order in which they arrive. Moreover, since a sequential order is followed, all processes eventually get executed and starvation is avoided. However, the turnaround time and the waiting time are found to be high. A scenario in which this algorithm is found to be least efficient is the one in which a long duration process is followed by a large number of shorter processes. All the short processes remain stuck in the queue, while the larger one executes [3].

II. **Shortest Job First (SJF)**

This is arguably one of the most efficient algorithms for a large number of processes with fixed priorities [2]. It requires advance knowledge of the burst times of all the processes, when they enter the job queue. Upon entering the ready queue, the processes with least burst time are placed near the front of the ready queue, whereas the ones with the longest burst times are placed near the rear of the queue. It is however, incredibly difficult to implement this algorithm as it is difficult to predict the burst times of the processes which will arrive next. There is a small chance of starvation, as there may be a large number of small processes, which might not allow a longer process to execute. In essence, this algorithm can be equated to a priority algorithm, which assigns highest priority to the shortest jobs.

The shortest job first helps in reducing the average waiting time of all the processes, however, this is usually at the cost of the average waiting time for processes with relatively higher burst times. The waiting time is reduced due to the fact that, while longer processes keep on getting shelved, shorter processes are immediately run and serviced. The only major liability of this algorithm is the total starvation of processes with higher service times [3].

III. **Round Robin (RR)**

The basic principle of round robin scheduling algorithm lies in its quantization of computing time. This quantization refers to the slicing of available computational time into smaller blocks of fixed sizes. The scheduler thus, travels across all the available processes and processes each of them for a fixed period of time in a sequential manner. New processes automatically get added to the back of the queue as and when they arrive [4]. The efficiency of Round Robin lies in the manner in which the quantum of time is decided. The longer the quantum becomes, the more likely it is to replicate the behavior of First Come First Serve (FCFS). The shorter it becomes, the least effective it will become as more time will be wasted in context switches. Moreover, due to high amount of context switching and due to the ability to process a single process for a fixed amount of time only, the ability to meet deadlines is lost.

IV. **Priority Scheduling (PS)**

The priority scheduling algorithm can be either static or dynamic. The static allocation of priorities can be on the basis of a predefined reference and results in priorities being allocated as soon as the processes arrive in the ready queue. The dynamic allocation of priorities results in the priorities of processes changing during the execution of the various processes [5]. The dynamic change of priorities can be referred to as the Shortest Job First (SJF), which results in priorities being calculated iteratively for all processes as soon as a process ends, which results in newer processes being assigned a priority. The other example of dynamic allocation is the Shortest Remaining Time First (SRTF) algorithm. This results in reevaluation of priorities after every time unit, which causes the priorities to be recalculated and CPU time to be assigned accordingly.

3 Comparison of Scheduling Algorithms—Gantt Chart, Waiting Time, Turnaround Time

The sample process table which we have chosen is as follows:

Process ID	Burst time (ms)
P1	10
P2	2
P3	8
P4	6

In accordance with the First Come First Serve algorithm, we get the following Gantt Chart in Fig. 2.

In accordance with the Shortest Job algorithm, we get the following Gantt Chart in Fig. 3.

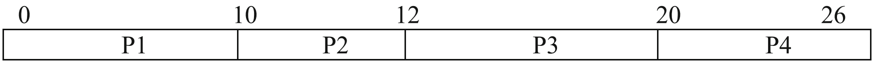


Fig. 2 Gantt chart for FFCS

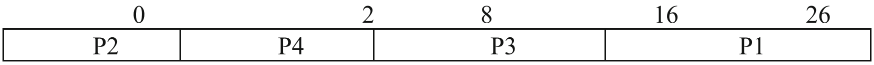


Fig. 3 Gantt chart for shortest job

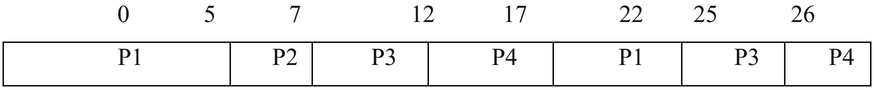


Fig. 4 Gantt chart for Round Robin

Table 1 Sample process table with burst time and priorities

Process ID	Burst time (ms)	Priority
P1	10	3
P2	2	1
P3	8	4
P4	6	2

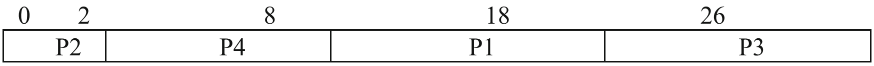


Fig. 5 Gantt chart for priority scheduling

In accordance with the Round Robin algorithm, we get the following Gantt Chart in Fig. 4.

For the simulations of the Priority Scheduling algorithm, we make use of priorities, which are assigned in following Table 1.

In accordance with the Priority Scheduling algorithm, we get the following Gantt Chart in Fig. 5.

The turnaround time for a process is calculated as the amount of time it takes for a particular process to be serviced from the moment it enters the ready queue, to when the process is fully serviced. The waiting time for a process is calculated as the amount of time a process spends in the ready queue from when it enters it to once the processing begins.

The above data in Tables 2 and 3 can be visualized in the form of a graph, which will allow us to analyze the efficiency of each type of scheduling algorithm in a better manner.

Table 2 Waiting time of individual processes for each scheduling algorithm

Process ID	Waiting time (ms)			
	FCFS	SJF	Round Robin	Priority scheduling
P1	0	16	12	8
P2	10	0	5	0
P3	12	8	17	18
P4	20	2	20	2
Average waiting time	10.5	6.5	13.5	7

Table 3 Turnaround time of individual processes for each scheduling algorithm

Process ID	Turnaround Time (ms)			
	FCFS	SJF	Round Robin	Priority scheduling
P1	10	26	22	18
P2	12	2	7	2
P3	20	16	25	26
P4	26	8	26	8
Average turnaround time	17	13	20	13.5

4 Simulation Results

It is clearly observed that the turnaround times for Shortest Job First (SJF) are significantly lower than the other scheduling algorithms [6, 7, 8]. While the First Come First Serve (FCFS) serves the best under batch processing, it is Shortest Job First (SJF) aim is to reduce both waiting time and turnaround time of jobs [6]. The Priority Scheduling (PS) serves a purpose only when a certain priority needs to be added to processes and they need to be processed only in accordance to them (Figs. 6 and 7).

5 Conclusion

The purpose of scheduling algorithm on computation environment is to schedule all type of jobs for reducing long waiting time at queue. In this paper, we have worked on different scheduling algorithm such as Round Robin (RR), Shortest Job First (SJF), First Come First Serve (FCFS) and Priority Scheduling (PS) in order to obtain comparative results. The SJF scheduling algorithm aim to server all shortest jobs first in result the processing time increases. Even SJF produces optimal job waiting time and average turnaround time but long process never get served. Every scheduling algorithm has its own advantages and disadvantages in terms of Processing Time and

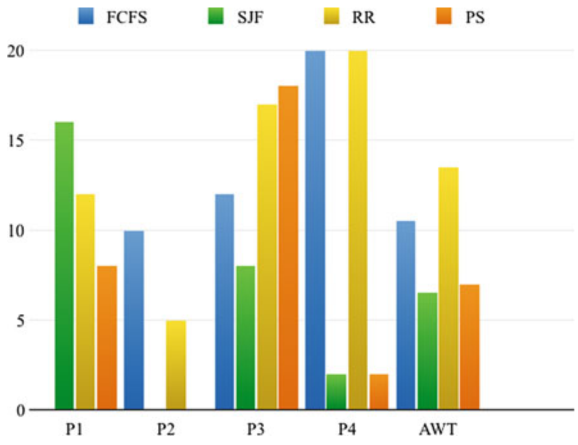


Fig. 6 Visualization of waiting times for various scheduling algorithms

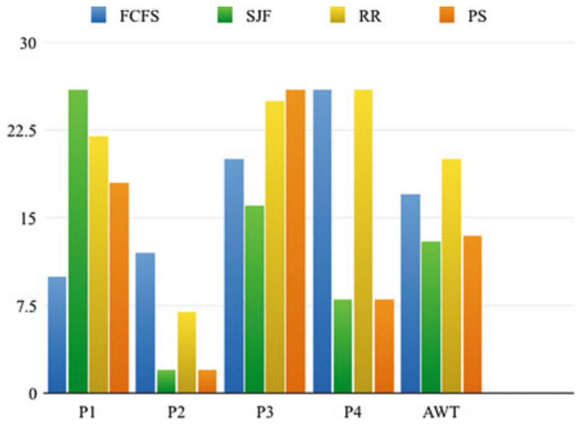


Fig. 7 Visualization of turnaround times for various scheduling algorithms

Turnaround Time. In order to evaluate performance of each scheduling algorithm is to code it and has to run on operating system Environment, this way we can evaluate the performance of the algorithm in real-time systems.

References

1. Sindhu, M., Rajkamal, R., Vigneshwaran, P.: An Optimum Multilevel CPU Scheduling Algorithm. 978-0-7695- 4058-0/10 $26.00 © 2010 IEEE
2. Silberschatz, A., Galvin, P.B., Gangne, G.: Operating System Concepts, 6 edn. Wiley, INC (2002)
3. Saleem, U., YounusJaved, M.: Simulation of CPU Scheduling Algorithm. 0-7803-6355-8/00/ $10.00 @ 2000 IEEE
4. Huajin, S., Deyuan, G., Shengbing, Z., Danghui, W.: Design Fast Round Robin Scheduler in FPGA. 0-7803-7547- 5/021/$17.00 @ 2002 IEEE

5. Mamunur Rashid, Md., NasimAdhtar, Md.: A new multilevel CPU scheduling algorithm. J. Appl. Sci. **6**(9), 2036–2039 (2009)
6. Suranauwarat, S.: A CPU Scheduling Algorithm Simulator. In: WI 37th ASEE/IEEE Frontiers in Education Conference, Milwaukee (2007)
7. Black, D.L.: Scheduling support for concurrency and parallelism in the mach operating system. Computer **23**, 123–126 (1985)
8. Bhardwas, A., Rachhpal Singh, Gaurav: Comparative study of scheduling algorithm in operating system. Int. J. Comput. Distrib. Syst. **3**(I) (2013)

Sentiment Analysis of Restaurant Reviews

R. Rajasekaran, Uma Kanumuri, M. Siddhardha Kumar, Somula Ramasubbareddy and S. Ashok

Abstract The project is on sentiment analysis on a data set retrieved from a data warehouse. For example, Amazon, one of the largest online shopping sites receives orders and reviews in millions every day. This huge amount of raw data can be used for industrial or business purpose by organizing according to our requirement and processing. This project provides a way of sentiment analysis using data mining techniques which will process the huge amount of product review data faster. We are going to work on this dataset of reviews and apply an algorithm to extract a meaningful result. The following literature survey illuminates the meaning of sentiment analysis, the various current methodologies of implementation, and the future scope in this domain.

1 Introduction

Currently, the amount of user comments, people's views and opinions broadcast to the world is growing exponentially. And there are many companies specializing in data extraction, data warehousing and selling information about market analysis and forecasting to the interested partied based on such user content being posted

R. Rajasekaran · U. Kanumuri · S. Ramasubbareddy (✉) · S. Ashok
Department of Computer Science and Engineering, VIT University, Vellore, Tamilnadu, India
e-mail: svramasubbareddy1219@gmail.com

R. Rajasekaran
e-mail: vitrajkumar@gmail.com

U. Kanumuri
e-mail: uma.kanumuri@gmail.com

S. Ashok
e-mail: sarabu.ashok@gmail.com

M. Siddhardha Kumar
Department of Information Technology, R.V.R & JC College of Engineering, Guntur, Andhra Pradesh, India
e-mail: siddardha1947@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_41

online. This project covers the sentiment analysis factor of user content and opinions found on the Internet. Users can express their ideas and share thoughts and views on various micro-blogging sites like Twitter and Word Press or a social media platform like Facebook and Google Plus. There are various online shopping websites that have added the feature to rate, comment, and share the product on sale. This helps the consumers to decide in the favor of buying that particular product. Sentiment analysis in this domain will help automate the highest voted ratings, the best-suited and positive ratings, and the most helpful ratings to be shown based on the numerous comments and product reviews given.

2 Background

This Bo Pang and Lee were the pioneers in this field [1]. Current works include mathematical expressions to evaluate sentences based on the proximity to adjectives and adverbs. Various mechanism has been implemented until now, which includes bags of words, training corpus, document level, sentence level, and feature-level opinion mining [2]. Different polarity measures exist according to the external system wherein sentimental analysis is utilized.

Sentiment analysis is a sub-field of artificial intelligence focused on text parsing and classifying it as positive, negative, or neutral. The Hadoop architecture consists of Common Utilities, Yam Framework, HDFS, and MapReduce paradigm [3]. HDFS (Hadoop Distributed File System) is a file system which is a distributed file system that runs on commodity machines [1]. It has high fault tolerance and is designed for low cost machines. HDFS has a high throughput access to application and is suitable for applications with large amount of data. They also perform block creation, deletion, and replication upon instruction. Replicating data pertaining to a file system increases the robustness of the system and adds to the integrity of data in the system. A programming model known as MapReduce is used for generation and processing of massive datasets. A map function is specified by the users that process a key/value pair in order to generate a set of intermediate key/value pairs, and a reduce function is specified which merges all the intermediate values that are associated with the similar intermediate keys. Programs which are implemented in this kind of functional style are automatically executed and parallelized on a cluster of machines. The details related to scheduling the execution of program across multiple machines, partitioning of input data, management of machine to machine communication and handling of failures is taken care by the run-time system.

There are several methods that have been implemented in this domain of research. Mane et al. [1] discuss their approach, which focuses on performance speed. They extract the Twitter tweets using a twitter API, which extracts unstructured data along with timestamp of every tweet. Next, POS (Part of Speech) tagging is performed, wherein the stop words are removed, unstructured tweet is converted into a structured one and then emoticons are evaluated, in this step. The words are then converted to root form so as to reduce the access time. All the root words have a sentiment value, based on the weighted mean of all its derived forms. Next MapReduce is applied on the Sentiment Directory based on the POS words and root words in the tweets. The accuracy of this method is around 72.27%, same as the accuracy of a Naïve Bayesian classifier.

While this method is safe and fast, there have been others too. Take for example an Infosys project by Dasgupta et al. [4]. They make use of Memory Spark RDD, Hadoop R, Map Reducer, Spark ML along with the on-demand computing options having minimal cost, reaping full benefits of the new paradigm in the usage of the technology for contemporary dimensions of advanced analytics, in this case analysis of sentiments. The authors [4] perform their analysis on Facebook comments to encapsulate the inputs and feedbacks on various brands. A Facebook access token was used, constructing a Uniform Resource Locator (URL) to the Graph API of Facebook using which the comments in Facebook are collected. Arrays in JSON of comments and posts were acquired as a response from the URL. Partitions of 25 posts a page were obtained as results. All of the comments and posts for a given search string, i.e., the brand was collected iteratively. Next, text cleaning had to be done to convert the strings into lower case and strip them of all http content, URLs, punctuations, special characters, and numbers. Stop words were removed in the subsequent step. All the articles and connector words in English provide little value to the analysis phase, and are hence called stop words. Next, N-gram generation step is done, where the number of times a series of words is repeated is calculated according to the value of "N". This acquired data is given as input to the HDFS which is then worked on by the RDD (Resilient Distributed Dataset). This file then becomes the input to the analysis program written in R. The code was written in R using the sentiment analysis package from CRAN. The polarity and category are determined using the algorithm and scores are assigned for every word in a sentence. The normalization of the score is done by dividing it by the total number of words found in its respective category. Later, the resultant's log value is determined. The process is then repeated for all the words that can be found in the sentence, and all positive and negative values are noted, and normalized. Finally, the average of the values is taken, and the sentiment is shown. The authors achieved a mean accuracy of 67.6% for six brands.

Sentiment analysis focuses on feature-based opinion summarization, which is very clearly explained by Mehta et al. [5]. It identifies the features in the given review and expresses the sentiment relevant to the feature. For, e.g., battery and heating are features of a handheld device. The authors use Apache Hadoop and its MapReduce to analyze the web scraped data, and send the result back to the website for other users to know. They implement two Map-Reduces—first for feature recognition and part of speech tagging and the second to carry out the sentiment analysis and calculate the overall value. First MapReduce has to carry out the steps of Sentence Detection, finding meaningful sentences in big paragraphs; Punctuation Removal, removing special characters and punctuations to clean the data; Phrase removal, vague and/or long negative phrases are replaced by a negation; removing stop words, feature category, wherein features are identified and classified, and finally part of speech tagging using Apache's OpenNLP. The second MapReduce makes use of words, especially adjectives and adverbs which affect the meaning of sentences. First step is Words classifier, in which all POS tagged adjectives and adverbs are searched for to get their value [6]. Next comes SentiWordNet values obtained using the SentiWordNet open source lexical resource. Finally the overall value is calculated. As their conclusion, the authors assert that the use of MapReduce and Hadoop architecture, the time complexity was greatly reduced.

There have been some works in this field using machine learning as well. Hybrid classification technique has been used for sentiment classification of movies reviews. Integration of different feature sets and classification algorithms such as Naïve Bayes, genetic algorithm has been carried out to analyze performance on the basis of accuracy. The output of research works shows that hybrid NB-GA is efficient and effective than base classifier and comparing in NB and GA, GA is more efficient than NB. If we consider text, mining, polarity of the document and words is also an important aspect. In the paper presented by Kawade and Oza [7], we can see that six different sentiments were be analyzed using sentiment package namely anger, disgust, fear, joy, sadness, and surprise. By using word cloud frequently occurring words were recorded. A sentiment was added to these frequently occurring words. These new words and sentiments are added to sentiment file for sentiment analysis. Present uses Bayes algorithm. Sentiment analysis algorithm compares each word with words in sentiment file and assigns count for each sentiment. Finally it can display count for each sentiment. This work also finds polarity of text. Polarity will be positive, negative or neutral. Herein, new words were identified using word cloud and then polarity was assigned to them. Similar to sentiment analysis, it also compares each word with polarity word file and counts polarity of text file. Lastly it displays count for each polarity. The authors extracted the tweets based on Uri attacks in Jammu and Kashmir from Twitter using Flume, and stored them in Hadoop's HDFS. Then, data preprocessing is carried out and the code was written in R. The authors made use of the tm, sentiment, and word cloud packages in R. As per the published results, 55.59% expressed fear about this event, and that polarity was determined to be around 67% negative, 15% neutral and 18% positive tweets by users.

Further research work is going on in various problem statements, such as developing a single big data platform for TV analysis that extracts views from TV social response in real time [8]. The authors have described the use of a SDN (software design network) to work with social media analytics system. Louis-Philippe Morency and his colleagues combine textual, acoustic, and video features in order to assess opinion polarity in 47 different videos on YouTube. They are yet to work in towards the fine-grained analysis phase [9].

3 Proposed Methodology

For our project, we propose to implement the Naïve Bayesian Classifier algorithm to classify reviews as positive or negative. A family of simple probabilistic classifiers in machine learning called Naive Bayes classifiers are based on the application of

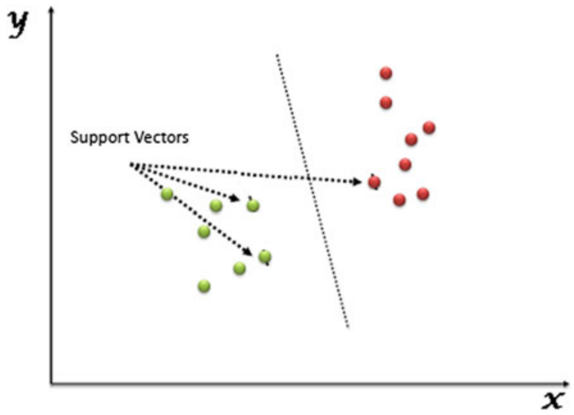


Fig. 1 Simple graph explaining SVM

Bayes' theorem with assumptions which are independent between the features. The mathematical theory behind Bayes has been explained with the help of the following few terms.

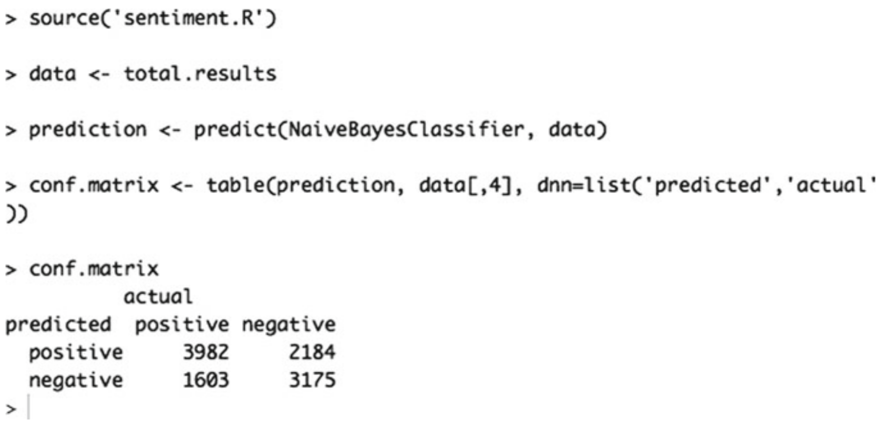
Theory:

$$
\begin{aligned}
&P(A|B) = \text{Fraction of worlds in which } B \text{ is true that also have } A \text{ true } P(A^{\wedge}B)\\
&P(A|B) = P(A^{\wedge}B)/P(B)\\
&\text{Corollary}: P(A^{\wedge}B) = P(A|B)P(B)\\
&P(A|B) + P(\ A|B) = 1\\
&\sum P(A = ve|B) = 1
\end{aligned}
$$

Observed data and prior knowledge can be combined using practical learning algorithms provided by Bayesian classification. A convenient perspective for evaluating and understanding learning algorithms is provided by Bayesian classification. Explicit probabilities can be calculated by Bayesian classification.

There is yet another method that we can use to implement sentiment analysis and that is Support Vector Machines (SVM). Supervised learning models like Support Vector Machines (SVM) are associated learning algorithms which are used for regression and classification analysis. When provided with data consisting of a set of training examples, each marked as belonging to a certain category, the Support Vector Machine algorithm classifies new examples to a category accordingly, thus making Support Vector Machine algorithm a non-probabilistic binary linear classifier in Fig. 2 (Fig. 1).

In Support Vector Machines, the learning of the hyperplane is done by transformation of the problem using linear algebra.



```
> source('sentiment.R')

> data <- total.results

> prediction <- predict(NaiveBayesClassifier, data)

> conf.matrix <- table(prediction, data[,4], dnn=list('predicted','actual'
))

> conf.matrix
          actual
predicted  positive negative
  positive     3982     2184
  negative     1603     3175
> 
```

Fig. 2 Confusion matrix

The equation of linear kernel for the prediction for a new input using the dot product between each support vector (*xi*) and the input (*x*) is calculated as follows.

$$f(x) = B(0) + \text{sum}(ai * (x, xi))$$

This here is an equation the involves calculation of the inner products of all support vectors in training data with a new input vector (*x*). The coefficients ai (for every input) and B0 must be estimated by the learning algorithm for the training data.

The polynomial kernel can be written as

$$K(x, xi) = 1 + \text{sum}(x * xi)^\wedge d \text{ and exponential as } K(x, xi)$$
$$= exp(-\text{gamma} * \text{sum}((x - xi^2))).$$

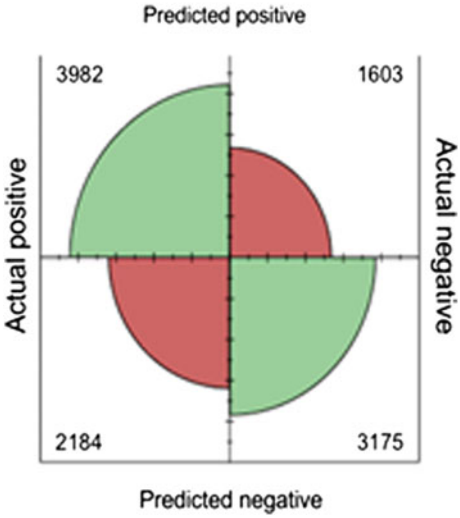
As a part of our results after applying Naïve Bayesian Classifier to our restaurant review data set, the following observations were made. The confusion matrix so obtained between the actual and predicted values of positive reviews and negative reviews is shown below. We can infer from it that the predictions are above average positive and provide a near correct feedback about the reviews.

Based on this confusion matrix in Fig. 2, we can also provide a graph to sustain our accuracy statement in Fig. 3. This clearly shows us what percentage of predictions was correct. The naïve Bayesian classifier has an accuracy of 72.2%.

4 Comparisons

Theoretically, it is very difficult to gauge the effectiveness and the difference in performance of both the aforementioned methods SVM and Naïve Bayesian. SVM

Fig. 3 Graph of actual versus predicted values



is the most useful to exploit the hidden relations between parameters, which the latter method is unable to detect. This is because SVM works geometrically, whereas Naïve Bayesian works on a probabilistic approach to the problem. But, in the end, it basically boils down to good performance and better results with the simplest solution possible. For example, spam detection has famously been solvable by simple Naïve Bayesian classifiers. Face recognition in images by a similar method enhanced with boosting is another example. Thus, we too use Naïve Bayesian Classifier as a suitable method to predict restaurant reviews through Twitter tweets.

5 Conclusion

As we have seen, there have been various innovations, and various reasons for research in this domain. We benefit a lot from this topic, both knowledge-wise and experience-wise due to its applicative nature. Sentiment analysis has been a hot topic for the past several years mainly because of its wide range of capabilities and diverse applications and huge untapped reserves, yet to be explored. Our project shall revolve around sentiment analysis of product reviews to provide accurate, concise and relevant information.

Feature Enhancement

Even though there have been various developments in the field of sentiment analysis and newer technologies are being developed as we write this, there are yet many areas to focus on. They are mentioned as follows:

Data collected from various sources is often noisy, unstructured and lacks uniformity. Data preprocessing takes the maximum processing and time in this whole process of SA.

Most dictionaries offer meanings but rarely do you find one describing the latest and trending acronyms and slangs in the world. Slangs are a part of the growing internet and should also be considered into SA process with equal zeal.

There is an obvious lack of universal grading for opinion across the sentiment dictionaries. Hence, there will always be discrepancies in evaluating and comparing the numerous methodologies.

Non-English language area is a mine of opportunities since this area has not been explored for NLP and opinion mining system.

Irony and Sarcasm are the heart and life of online discussion portals. Applying traditional SA methods to sarcasms return totally opposite answers. Very less study has been devoted in this regard. There is a need for the development of more computational approaches based on the appraisal theory.

Almost no research has been done for multimodal analysis, i.e., combining audio, video and text sentiment analysis together. These are just a few and there is much work to be done yet.

References

1. Mane, S.B., Sawant, Y., Kazi, S., Shinde, V.: Real time sentiment analysis of twitter data using hadoop. Int. J. Comput. Sci. Inf. Technol. (IJCSIT) **5**(3), 3098–3100 (2014)
2. Mishra, N., Jha, C.K.: Classification of opinion mining techniques. Int. J. Comput. Appl. **56**(13) (2012)
3. Rodrigues, A.P., Chiplunkar, N.N., Rao, A.: Sentiment analysis of social media data using Hadoop framework: a survey. Int. J. Comput. Appl. **151**(6) (2016)
4. Dasgupta, S.S., Natarajan, S., Kaipa, K.K., Bhattacherjee, S.K., Viswanathan, A.: Sentiment analysis of Facebook data using Hadoop based open source technologies. In: IEEE International conference on data science and advanced analytics (DSAA), 2015. 36678 2015, pp. 1–3. IEEE, Oct 2015
5. Mehta, J., Patil, J., Patil, R., Somani, M., Varma, S.: Sentiment analysis on product reviews using Hadoop. Int. J. Comput. Appl. **142**(11) (2016)
6. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Commun. ACM **51**(1), 107–113 (2008)
7. Kawade, D.R., Oza, K.S.: Sentiment analysis: machine learning approach
8. Hu, H., Wen, Y., Gao, Y., Chua, T.S., Li, X.: Toward an SDN-enabled big data platform for social TV analytics. IEEE Netw. **29**(5), 43–49 (2015)
9. Cambria, E., Schuller, B., Xia, Y., Havasi, C.: New avenues in opinion mining and sentiment analysis. IEEE Intell. Syst. **28**(2), 15–21 (2013)

Content-Based Movie Recommendation System Using Genre Correlation

SRS Reddy, Sravani Nalluri, Subramanyam Kunisetti, S. Ashok and B. Venkatesh

Abstract A recommendation system is a system that provides suggestions to users for certain resources like books, movies, songs, etc., based on some data set. Movie recommendation systems usually predict what movies a user will like based on the attributes present in previously liked movies. Such recommendation systems are beneficial for organizations that collect data from large amounts of customers, and wish to effectively provide the best suggestions possible. A lot of factors can be considered while designing a movie recommendation system like the genre of the movie, actors present in it or even the director of the movie. The systems can recommend movies based on one or a combination of two or more attributes. In this paper, the recommendation system has been built on the type of genres that the user might prefer to watch. The approach adopted to do so is content-based filtering using genre correlation. The dataset used for the system is Movie Lens dataset. The data analysis tool used is R.

1 Introduction

In this age of the Internet, the quantity of data transactions that happen every minute has increased exponentially. The huge amount of data has dramatically increased with the number of users on the Internet. However, not all the data available on the Internet is of use or provides satisfactory results to the users. Data in such huge volumes often turns out to be inconsistent and without proper processing of this information, it gets wasted. In such cases, users have to run their search multiple times before they finally obtain what they were originally looking for. To solve this problem, researchers have come up with recommendation systems. A recommendation system provides relevant information to the users by taking into account their past preferences. Data is filtered and personally customized as per the user requirements. With more and more data available on the Internet, recommendation systems

S. Reddy (✉) · S. Nalluri · S. Kunisetti · S. Ashok · B. Venkatesh
VIT University, Vellore, Tamilnadu, India
e-mail: janani333333@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_42

have become really popular, due to their effectiveness in providing information in a short time-span. Recommender systems have been developed in various areas such as music, movies, news, and products in general. In today's age, a majority of organizations implement recommendation systems for fulfilling customer requirements. LinkedIn, Amazon, and Netflix are just a few to name. LinkedIn recommends relevant connections of the people the user might know among the millions that are subscribed on the portal. This way, the user does not have to run extensive searches for people manually. Amazon recommendation systems work such that they suggest correlated items that the customers can purchase. If a certain customer prefers buying books from the shopping portal, Amazon provides suggestions related to any new arrivals in previously preferred categories. In a very similar way, Netflix takes into account the types of shows that a customer watches, and provides recommendations similar to those. By the method in which recommendation systems work, they can be broadly classified into three categories—Content-based, Collaborative and Hybrid approach. A content-based recommendation system considers the user's past behavior and identifies patterns in them to recommend to recommend items that are similar to them. Collaborative filtering analyses the user's previous experiences and ratings and correlates it with other users. Based on the ones that have the most similarity, recommendations are made. Both content-based- and collaborative-based filtering have their own limitations. To overcome this, researchers suggested a hybrid approach which would combine the advantages of both the methods. This paper suggests a content-based recommendation system that utilizes genre correlation. The dataset used for this purpose is a Movie Lens dataset containing 9126 movies which are classified according to genres. There are a total of 11 genres. The ratings for these moves have been collected from 671 users. By taking into account the movies which received high ratings from the users, movies containing similar genres are recommended to them.

2 Background

Recommender systems are broadly classified into three types—collaborative filtering systems, content-based filtering systems, and hybrid systems [3]. Collaborative systems utilize inputs from various users and run various comparisons on these inputs [3]. They build models from the past behavior of the users [1]. Movie recommendation systems, for example, utilize the ratings of users for various movies [2], and attempt to find other like-minded users, and recommend movies they have rated well [3]. Collaborative filtering systems have two approaches—memory-based approaches and model-based approaches [3]. Memory-based approaches continuously analyze user data in order to make recommendations [3]. As they utilize the user ratings, they gradually improve in accuracy over time [3]. They are domain-independent and do not require content analysis [3]. Model-based approaches develop a model of a user's behavior and then use certain parameters to predict future behavior [3]. The use of partitioning-based algorithms also leads to better scalability and accuracy [3].

Content-based filtering systems analyze documents or preferences given by a particular user, and attempt to build a model around this data [3]. They make use of a user's particular interests and attempt to match a user's profile to the attributes possessed by the various content objects to be recommended [3]. They have the added disadvantage of requiring enough data to build a reliable classifier [1]. Content-based filtering systems are divided into three methods—wrapper methods, filter methods, and embedded methods [3]. Wrapper methods divide the features into subsets, run analysis on these subsets and then evaluate which of these subsets seems the most promising [3]. Filter methods use heuristic methods to rate features on their content [3]. Both these methods are independent of the algorithms used. In contrast, embedded methods are coupled with the algorithm used—feature selection is performed during the training phase [3]. Hybrid systems combine collaborative and content-based filtering systems, in order to optimize the recommender systems, and reduce the drawbacks present in each of the two methods [3]. Thus, it tries to stretch the benefits of one method to compensate for the disadvantages of the other [3]. There are three types of hybrid systems—weighted hybrid, mixed hybrid, and cross-source hybrid [3]. In weighted hybrid systems, a score is maintained for each object, finding the weighted sum with respect to the various context sources [3]. These are given different weights based on a user's preferences [3]. In mixed hybrid approaches, each source is used to rank the various items, and the top few items from each rank list are picked [3]. Cross-source hybrid methods recommend items that appear in multiple context sources [3]. These methods work on the principle that the more sources an item appears in, the more important the item [3]. Wakil et al. attempted to improve their recommendation system by filtering using emotions [4]. When a user watches a certain type of movie, certain emotions are triggered from within them [4]. In the same way, the emotions of a user can trigger the need to watch a certain type of movie [4]. They recognized that traditional user profiles do not take into account the user's emotional status, and designed an algorithm that utilizes emotion determination [4]. It analyses a color sequence chosen by the user in accordance with his emotions to determine current emotional state of the user [4]. Debnath et al. proposed a hybrid recommendation system that utilizes feature weighting [5]. They determined the importance of various features to each user, and accordingly assigned weights to these features [5]. They then found the weighted sum in order to predict which items would further interest the user [5].

3 Recommendation System Using Content-Based Filtering

The approach used for building the recommendation system is content-based filtering. As discussed earlier, content-based filtering analyses user's past behavior and recommends items similar to it based on the parameters considered. This aims at recommending movies to users based on similarity of genres. If a user has rated high for a certain movie, other movies containing similar genres are recommended by the system. The dataset used in for this purpose is subdivided into two sections.

One section contains the list of movies along with the genres that they have been categorized under. The other part of the dataset contains a list of ratings of movies that have been rated by the user on a scale of 1–5, with 5 being the highest. First, a combined dataset of movies, genres and their ratings has to be constructed for correlating genres with the ratings. For the sake of simplicity, the ratings have been converted to binary values. If the rating given by a particular user is greater than 3, it receives a value of 1, otherwise it receives a value of -1. The genres are also segregated in a binary format, maintaining a consistent approach. Out of the set of 11 genres that are present in total, if a movie has a certain genre, it receives the value of 1. If the genre is not present in the movie, it receives a value of 0. The user profile matrix provides a combined effect of the genres and ratings by computing the dot product of the genre and the ratings matrix. Again for the sake of consistency, a binary format is adopted. If the dot product is a negative value, 0 is assigned to it. For a positive value, 1 is assigned to it. After obtaining a dot product matrix of all the movies, a similarity measure is calculated by computing the least distance between the user under consideration and the others. The values which have the least deviation with respect to the current user's preferences are the ones that are recommended by the system. The algorithm adopted for building the recommendation system is given below:

Algorithm Step 1. Construct a data frame of the genre dataset with movie ID as the rows and genres as columns separated by pipeline character.
Step 2. Make a list of all the genres that are available in the dataset.
Step 3. Iterate through the previously made genre data frame. If a genre is present in a movie, value of 1 is assigned to the genre matrix.
Step 4. Read the ratings sheet and construct a ratings matrix which assigns 1 for movies which has rating more than 3 and -1 for movies which has ratings less than or equal to 3.
Step 5. Calculate the dot product of the two matrices—genre matrix and ratings matrix. This is the result matrix
Step 6. Convert the result matrix to a binary format. For a negative dot product value, assign 0, else assign a value of 1.
Step 7. Calculate the Euclidian distance between the current user and other users.
Step 8. Retain the rows which have the minimum distance. These are the recommended movies for the current user.

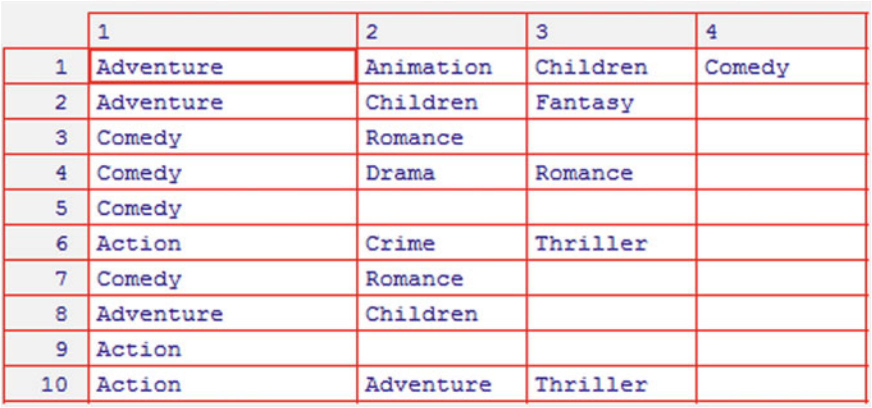
Offloading is a method of transferring resource-intensive application from portable device to remote server by considering different parameters. Offloading mechanisms involves three tasks before it get executed. They are partitioning, profiling, offloading decision.

4 Simulation Results

The genre matrix constructed with rows containing movies and genres separated by columns. There are a total of 11 genres in the dataset (Fig. 1).

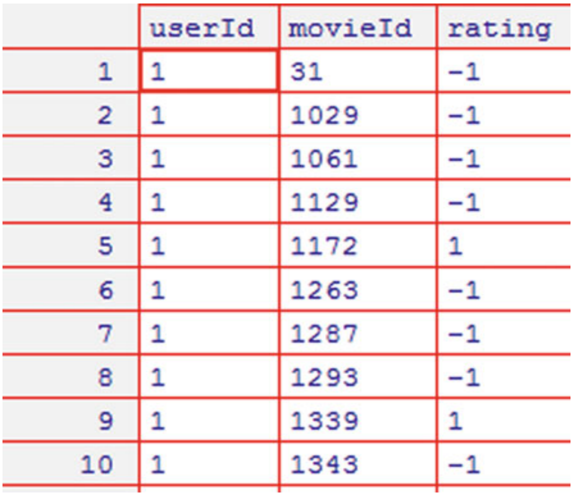
The ratings matrix for each user corresponding to the movie ID is converted to a binary format. Every user has rated one or more than one movie (Fig. 2).

Using the genres matrix and ratings matrix, the result matrix is computed which is the dot product of the previous two matrices. The result is further converted in a binary format in Fig. 3. If the value of the dot product is more than 0, 1 is assigned to that cell otherwise 0 is assigned.



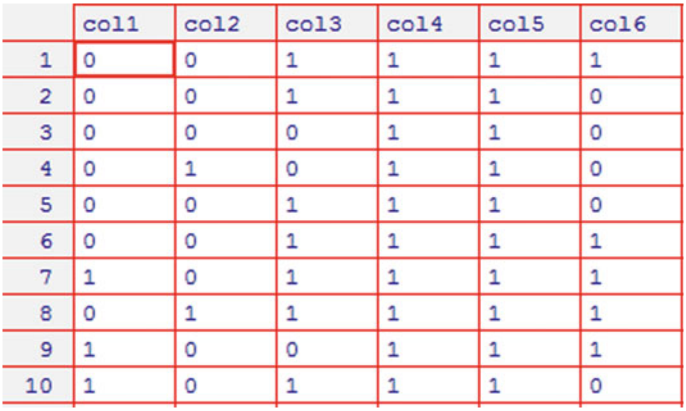
	1	2	3	4
1	Adventure	Animation	Children	Comedy
2	Adventure	Children	Fantasy	
3	Comedy	Romance		
4	Comedy	Drama	Romance	
5	Comedy			
6	Action	Crime	Thriller	
7	Comedy	Romance		
8	Adventure	Children		
9	Action			
10	Action	Adventure	Thriller	

Fig. 1 Genre matrix



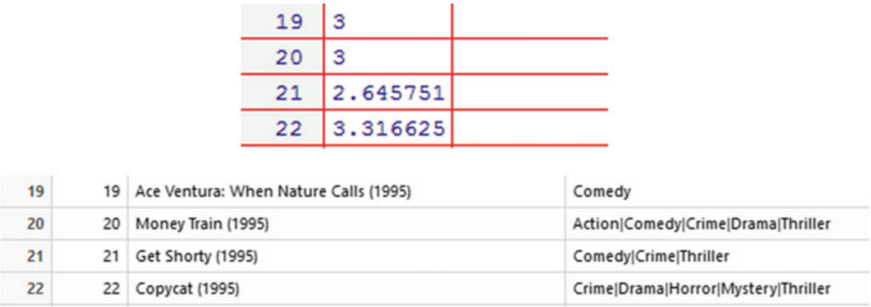
	userId	movieId	rating
1	1	31	-1
2	1	1029	-1
3	1	1061	-1
4	1	1129	-1
5	1	1172	1
6	1	1263	-1
7	1	1287	-1
8	1	1293	-1
9	1	1339	1
10	1	1343	-1

Fig. 2 Ratings matrix



	col1	col2	col3	col4	col5	col6
1	0	0	1	1	1	1
2	0	0	1	1	1	0
3	0	0	0	1	1	0
4	0	1	0	1	1	0
5	0	0	1	1	1	0
6	0	0	1	1	1	1
7	1	0	1	1	1	1
8	0	1	1	1	1	1
9	1	0	0	1	1	1
10	1	0	1	1	1	0

Fig. 3 Result matrix



19	3	
20	3	
21	2.645751	
22	3.316625	

19	19	Ace Ventura: When Nature Calls (1995)	Comedy
20	20	Money Train (1995)	Action|Comedy|Crime|Drama|Thriller
21	21	Get Shorty (1995)	Comedy|Crime|Thriller
22	22	Copycat (1995)	Crime|Drama|Horror|Mystery|Thriller

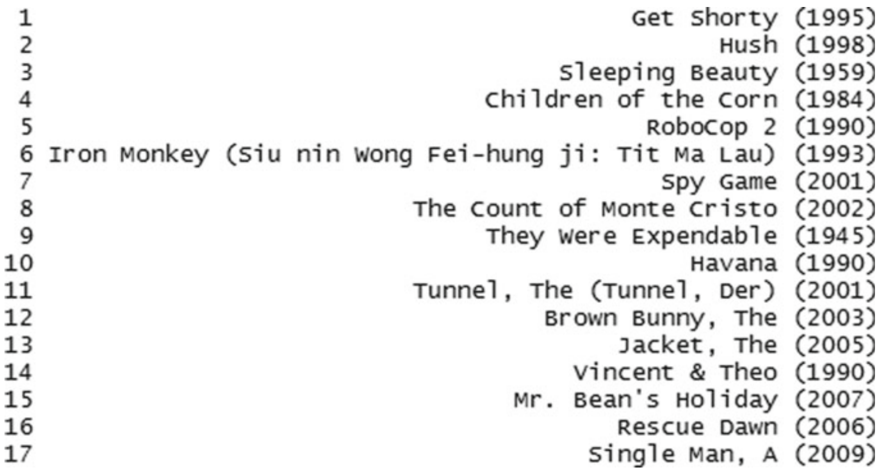
Fig. 4 Euclidean distance

After computing the result matrix, the Euclidean distances with respect to the other users are obtained and the ones having the minimum value is recommended as represented in Fig. 4.

Figures 5 and 6 shows the output of the various movies that have been recommended to the users based on their previous behavioral patterns.

5 Conclusion and Future Work

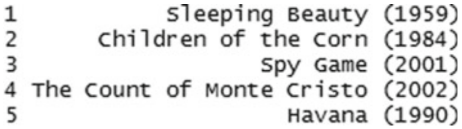
The recommendation system implemented in this paper aims at providing movie recommendation based on the genres of the movies. If a user highly rates a movie of a particular genre, movies containing similar genres will be recommended to him. Recommendation systems are widely used in today's era of Web 2.0 for searching for reliable and relevant information. While simple recommendation systems recommend users based on a few parameters, complex ones take many parameters into consideration. By implementing machine learning in recommender systems, intel-



```
1                                                   Get Shorty (1995)
2                                                         Hush (1998)
3                                              Sleeping Beauty (1959)
4                                         Children of the Corn (1984)
5                                                    RoboCop 2 (1990)
6 Iron Monkey (Siu nin Wong Fei-hung ji: Tit Ma Lau) (1993)
7                                                     Spy Game (2001)
8                                    The Count of Monte Cristo (2002)
9                                        They Were Expendable (1945)
10                                                      Havana (1990)
11                                   Tunnel, The (Tunnel, Der) (2001)
12                                            Brown Bunny, The (2003)
13                                                 Jacket, The (2005)
14                                              Vincent & Theo (1990)
15                                          Mr. Bean's Holiday (2007)
16                                                 Rescue Dawn (2006)
17                                               Single Man, A (2009)
```

Fig. 5 User 1 recommendations

Fig. 6 User 2 recommendations



```
1           Sleeping Beauty (1959)
2      Children of the Corn (1984)
3                  Spy Game (2001)
4 The Count of Monte Cristo (2002)
5                    Havana (1990)
```

ligent recommendations can be made for customers. Given the potential of such systems, they have a huge commercial value. Several MNCs have been exploiting the potential of recommendation system to lure customers into using their products. This also impacts greatly on the field of data mining and web mining.

Mobile cloud computing (mcc) is able to save energy, improve application and experience of the users. All frameworks mentioned above have their own benefits and issues but still not up to level to address all issues related to security, energy and user experience. Security issues are key problem in mcc, they need to be focused more compare to other issues.

References

1. Ghuli, P., Ghosh, A., Shettar, R.: A collaborative filtering recommendation engine in a distributed environment. In: 2014 International Conference on Contemporary Computing and Informatics (IC3I). IEEE (2014)
2. Zhao, L., et al.: Matrix factorization + for movie recommendation. In: IJCAI (2016)
3. Bhatt, B.: A review paper on machine learning based recommendation system. Int. J. Eng. Dev. Res. (2014)
4. Wakil, K., et al.: Improving web movie recommender system based on emotions. (IJACSA) Int. J. Adv. Comput. Sci. Appl. **6**(2) (2015)
5. Debnath, S., Ganguly, N., Mitra, P.: Feature weighting in content based recommendation system using social network analysis. In: Proceedings of the 17th International Conference on World Wide Web. ACM (2008)

Iterative Approach for Frequent Set Mining Using Hadoop Over Cloud Environment

S. Prasanna, Subashini Narayan, M. K. NallaKaruppan, Chunduru Anilkumar and Somula Ramasubbareddy

Abstract Cloud computing initially gained popularity as it offered an alternative for handling the ever-growing size of data. One of the main advantages of Cloud computing is parallel processing of data, which causes the effect of pooling the resources of various systems. The proposed project aims to implement the feature for the purpose of data mining and will use the Apriori Algorithm to demonstrate the results. Hadoop platform will be utilized for this project. The system will receive a dataset and redistribute it to the nodes of the cloud. Here, Apriori algorithm will be applied upon the sections of the dataset and the results will then be combined to obtain the frequent itemsets in the global data. Using the frequent item sets, rule mining will be achieved.

1 Introduction

Mobile—The main driving force behind Cloud computing are the IT giants such as IBM, Amazon, Microsoft, etc. With the growing popularity of digitization given the benefits it offers in maintaining records, big data is common in more and more organizations. Often the size of data makes it impossible for a human to remove the noise in the data and recognize complex patterns and relations. Data mining is a combination of statistical sampling, estimation, hypothesis testing, and pattern recognition. In abroad sense, data mining is often referred to as knowledge discovery in database or KDD. The importance of data mining algorithms and improvement of the algorithms lies in the fact that with the ever-growing size of data, traditional methods of finding patterns or making predictions continue to grow obsolete. This is mainly because the algorithm has to process a huge amount of data that is also being constantly updated. Data mining algorithms focus on reducing the time and computing power required to process huge data sets. The problem of increasing requirement of computing power is solved using parallel processing. Naturally, data

S. Prasanna · S. Narayan · M. K. NallaKaruppan · C. Anilkumar · S. Ramasubbareddy (✉)
VIT University, Vellore, Tamil Nadu, India
e-mail: svramasubbareddy1219@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_43

mining algorithms have to be tweaked so that they can effectively distribute work load amongst various nodes and work efficiently [1, 2]. Most computers in a network never utilize their processing capability to the maximum extent. Cloud computing can be used to efficiently distribute big data and process it at the nodes. This means that the entire processing power need not be centralized and that instead of investing in a new high end system, the systems that are already in use can be utilized. Algorithms must be periodically revised and improved upon to better utilize the power of parallel computing [3, 4]. Cloud computing has given rise to a new business model that allows distributed storage and processing of data thus reducing the cost of IT infrastructure [5]. The usage of "clouds" for utility-based billing means that an organization can reduce costs depending upon the efficiency of algorithms deployed by it. Map/Reduce architecture forms the core operation for parallel computing using Hadoop. The map operation allows big data to be mapped to keys and then distributed amongst the nodes in cloud while the reduce operation gathers results from the nodes and combines them together. MapReduce has been used for the implementation of machine learning based methods such as classification, logic regression, linear support vector, etc.

2 Background

Big Data is an expression used to denote an enormous volume of data that is so massive that it is tough to process utilizing conventional computing strategies. In most business situations the size of data is too huge or it surpasses current handling limits. Big Data enables organizations to enhance operations and make quicker, more insightful decisions. The data is gathered from various sources including messages, cell phones, applications-mails, databases, etc. This information can enable an organization to increase valuable understanding to enhance incomes and growth levels, strengthen operations, improve Customer Feedback, and retain Customers. Big Data usually includes utilizing NoSQL and Distributed Computing techniques to scrutinize the data. Cloud Computing provides IT assets like Infrastructure, Platform, Software, and Storage as Services. Some of the trademark characteristics of Cloud Computing include: Adaptable Scaling, Availability of scalable provisioning, High Availability, Asset Pooling, On-demand and pay as per requirement and demand.

In simple terms, Big data can be described as the input and Cloud Computing can be described as the framework in which operations and jobs can performed on the Big Data for its analysis. Big Data can be organized and processed in a Cloud Environment. MapReduce was developed by a few employees of Google. MapReduce is a programming paradigm utilized for parallel computations. Google's version of MapReduce is likewise called MapReduce and it not accessible to the general population, but the key ideas and concepts were published. Apache Hadoop is a free and open source implementation of the MapReduce model. It is reliant upon its intrinsic Hadoop Distributed File System (HDFS). It is based on Java. The principal thought behind MapReduce is to divide the problem into two subproblems: a guide and a decrease stage. First, the input information gets cut into littler pieces and each lump is given to a machine that executes a mapper. The mapper forms that lump and gives key esteem matches as the yield. Than mapreduce structure gathers these sets, sorts them in view of the key and passes them to reducers. A reducer gets a key and a rundown of qualities that has a place that key and gives the outcomes. The principle thought behind MapReduce is to divide the proposition into two subproblems: Map phase and Reduce Phase. First, the input information gets cut into little pieces and each lump is given to a machine that executes a mapper. The mapper forms that lump and gives key-value pairs as the output. Than MapReduce structure gathers these sets, sorts them in view of the key and passes them to the Reducers. A reducer then performs the requisite computations and consequently outputs the result. The tasks isolated by the principal application are right off the bat processed initially by the map jobs in a totally parallel fashion [6]. The MapReduce system sorts the yields of the maps, which are then used as input to the reduce jobs. The Hadoop Distributed File System then stores both the input and the output of the job. Hadoop Distributed File System (HDFS) is a framework that holds an expansive amount of data in terabytes or petabytes and gives quick and versatile access to this data [7]. It stores records in a redundant manner over many nodes to give adaptation to resistance failure and high accessibility amid execution of parallel applications. HDFS breaks a document into blocks of constant size (default square size is 64 MB) to store over many nodes. Hadoop utilizes a NameNode as master node and various DataNodes as slave nodes. The NameNode allots ids to the data blocks and stores additional data related to it such as permission, name, etc., in the form of metadata thus enabling quick access to this data. Data Nodes are the individual nodes which store and recover the redundant blocks of different records.

The following diagram illustrates the basic architecture and relationship between various nodes in a Hadoop framework (Fig. 1).

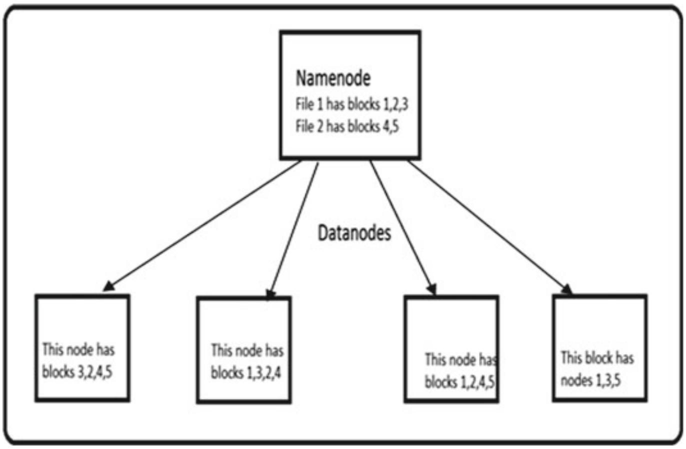

Namenode
File 1 has blocks 1,2,3
File 2 has blocks 4,5
Datanodes
This node has blocks 3,2,4,5
This node has blocks 1,3,2,4
This node has blocks 1,2,4,5
This block has nodes 1,3,5

Fig. 1 Data nodes and slave nodes in HDFS with block replication

Association Rule Mining endeavors to discover frequent itemsets among huge datasets and depicts the association relationship among various attributes. It was initially described in the paper "Fast Algorithms for Mining Association Rules" By R. Agarwal and R. Srikant. Apriori Algorithm is arguably the most popular algorithm used for Association Rule Mining. Apriori Algorithm utilizes an iterative approach for finding the frequent itemsets of a given dataset. It performs the search in an iterative manner layer-by-layer. It basically generates $k + 1$ itemsets by utilizing the k-itemsets. The basic algorithm can be described as

1. Let C_k and L_k be defined as the Candidate Itemset and Frequent itemset of "k" size respectively.
2. Initialize L_1
3. Now generate C_{k+1} from L_k.
4. For every transaction in the dataset, increase the count of all candidates in C_{k+1} that are included in that transaction.
5. Now generate L_{k+1} using the item candidates that have support greater than the minimum support.
6. Loop till L_k is not empty.
7. From this obtain all the k-sized frequent itemsets L_k.

The benefits of using this simple approach to Apriori algorithm is that is easy, intuitive, straightforward, no convoluted derivations are involved. Furthermore, the nature of this simple algorithm significantly diminishes the number of candidates to be verified and hence we can observe an enhancement in the efficiency of this algorithm.

Some of the disadvantages and overheads associated with this approach are

1. Colossal overhead is caused due to scanning the database multiple times. For datasets containing N-length frequent itemsets as the largest frequent itemset, we need to scan the database N number of times. This puts additional burden on the processor and hence slows down the entire process
2. It may create countless itemsets. Sometimes this number becomes very large and the algorithm becomes infeasible to implement.

Hence by generating large number of itemsets and by scanning the database multiple times, Apriori algorithm becomes computationally very expensive and will consume a lot of time as well as memory space. To overcome these limitations of the basic Apriori Algorithm, we utilize improved Apriori Algorithm. As a cloud computing environment can support parallel/distributed approach to computation, it can be utilized to optimize the existing Apriori Algorithm. Hence, parallel association rule mining techniques can be used for this purpose.

The improved algorithm can be described as

1. Divide the database into "n" subsets and distributed to "m" nodes for performing the parallel scan of the database.
2. Scanning of individual divided data set is performed at each node and then candidate set C_p is generated on the basis of this.

3. Support count of these candidate sets C_p is set to 1. After this, the candidate itemset is partitioned into and sent to “r” nodes along with their support counts.
4. These nodes then sum up the support counts of the same itemsets and to generate the final support count.
5. Establish the frequent itemset in the partition after tallying with the minimum support count as per the required application.
6. Merge the outputs from these nodes to produce the final global frequent itemset.

The aforementioned improved Apriori algorithm is utilized to extensively lessen the time as in this calculation the database needs to be scanned only once instead of the multiple times in the original Apriori Algorithm.

This modified and improved Apriori Algorithm can hence be implemented on Apache Hadoop Framework with the MapReduce Model as this Framework has the requisite tools and methodologies to successfully implement this modification. The algorithm can then be described as [8, 9, 10]

(1) The database is partitioned into n sets and are then sent to m nodes for processing by executing Map jobs.
(2) The datasets need to be in the form of a key and value pair for processing in MapReduce Model. Hence, this formatting needs to be done.
(3) Candidate sets are generating by the mapping function running in each node.
(4) The results are then combined by the combiner function.

The drawbacks and benefits of using this MapReduce

Advantages	Disadvantages
Reduction of Network Bandwidth Usage	Can only operate on data which is in the format of key, value pairs
Parallelization and efficient workload distribution	Next stage of the computation cannot be started until and unless the reducer has finished executing
Makes the entire problem very scalable	The data distribution version of Apriori is not suitable to be implemented using MapReduce
A platform which enables both distributed storage and high computing power	

3 Proposed Methodology

Using the improvements provided by Lian et al. [4] we intend that the algorithm reads through the entire dataset only once. Additionally, with each subsequent iteration, to find the larger frequent itemset, the data to be considered is reduced. For this purpose, we utilize the core operations of Hadoop that are MapReduce. While mapping, as the transactional dataset is scanned, we assign each item as a key and the value as one. Once the mapping procedure is complete, the Hadoop framework distributes data to all available nodes. For reduce operation, the reducer counts the occurrence of each item, combines the findings and eliminates the items that do not reach the minimum support. The generation of k-itemsets involves referring to the $k-1$ itemsets where the itemset is key and the number of occurrence is value. The frequent $k-1$ itemsets that hold similar items are the keys and the corresponding values will be sent to the same reducer job. Subsequently, the same reducer procedure is repeated where an itemset of size k is generated by combining the available $k-1$ itemsets and their count is compared to the minimum support count. The values corresponding to $k-1$ itemsets are further used in parallel for rule mining. The key-value combination is saved in corresponding files or other available memory. This means that during association rule mining, the key-count of both k-itemsets and $k-1$ itemsets are already available and need not be computed again. This saves processing time as well as provides the association rules for all frequent itemsets from size two to k, where k is the size of the largest itemset that satisfies the minimum support count.

4 Simulation Results

The running times of the both the original and improved algorithms were plotted against the no. of jobs to execute. The dataset chosen was a data set containing 1000 transactions. The running times for each of these situations were measured independently five times and the result chosen for plotting this graph is the average of these five observations. The results observed could be tabulated in the form of a graph as shown (Fig. 2).

In the above graph, the y-axis represents the seconds required to execute the MapReduce job and the x-axis represents the no. of jobs (the no. of cores/processors). As we can see, increasing the degree of concurrency to six or eight jobs results in no significant performance improvement and reaches almost saturation levels w.r.t reduction in processing time and speed of execution. This can be attributed to the fact that this testing was carried out in an environment where the CPU had four cores and increasing the degree of concurrency would not increase the utilization of CPU

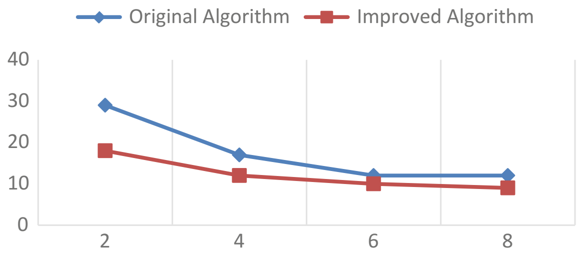


Fig. 2 Comparison of running times of the algorithms plotted against the number of jobs to execute

further. Also, as a general trend, it can be concluded from the graph that execution time for the improved algorithm is much faster than the running time of the original algorithm when CPU utilization is high and the CPU is not overwhelmed with a high no. of concurrent jobs.

5 Conclusion

In this paper, we discussed and analyzed on how to improve the performance and efficiency of frequent set mining using Apriori algorithm. For this purpose, a cloud-based environment utilizing HadoopMapReduce framework was proposed in this paper. Furthermore, an improved Apriori algorithm was proposed in this paper that further optimizes the performance of the MapReduce job by using the intermediate results obtained instead of scanning the original data set repeatedly resulting in massive overheads. The performance of these two algorithms was compared by comparing the running times of the algorithms in different scenarios. As a general rule of thumb, it could be observed from the tabulated results that the improved algorithm performs significantly faster than the original Apriori algorithm. Hence, an improved and robust apriori algorithm technique was put forward in this paper for frequent rule mining in a cloud environment.

References

1. Ekanayake, J., Fox, G.: High performance parallel computing with clouds and cloud technologies. In: International Conference on Cloud Computing. Springer, Berlin, Heidelberg (2009)
2. Sheth, N.R., Shah, J.S.: Implementing parallel data mining algorithm on high performance data cloud. Int. J. Adv. Res. Comput. Sci. Electr. Eng. (IJARCSEE) **1**(3), 45 (2012)
3. Jin, R., Yang, G., Agrawal, G.: Shared memory parallelization of data mining algorithms: techniques, programming interface, and performance. IEEE Trans. Knowl. Data Eng. **17**(1), 71–89 (2005)
4. Lian, W., et al.: Cloud computing environments parallel data mining policy research. Int. J. Grid Distrib. Comput. **8**(4), 135–144 (2015)
5. Chang, X.-Z.: Mapreduce-Apriori algorithm under cloud computing environment. In: 2015 International Conference on Machine Learning and Cybernetics (ICMLC), vol. 2. IEEE (2015)
6. Ezhilvathani, A., Raja, K.: Implementation of parallel apriori algorithm on hadoop cluster. Int. J. Comput. Sci. Mob. Comput. **2**(4), 513–516 (2013)
7. Tiwary, M., Sahoo, A.K., Misra, R.: Efficient implementation of apriori algorithm on HDFS using GPU. In: 2014 International Conference on High Performance Computing and Applications (ICHPCA). IEEE (2014)
8. Vajk, I.A.: Performance evaluation of Apriori Algorithm on a Hadoop cluster. Wseas. Us, pp. 114–121 (2013)
9. Singh, S., Garg, R., Mishra, P.K.: Review of apriori based algorithms on mapreduce framework. arXiv preprint arXiv:1702.06284 (2017)
10. Saabith, A.L.S., Sundararajan, E., Bakar, A.A.: Parallel implementation of apriori algorithms on the hadoop-mapreduce platform-an evaluation of literature. J. Theor. Appl. Inf. Technol. **85**(3), 321 (2016)

A Case Study on Recommendation Systems Based on Big Data

M. Sandeep Kumar and J. Prabhu

Abstract Recommender systems mainly utilize for finding and recover contents from large datasets; it has been determining and analysis based on the scenario—Big Data. In this paper, we describe the process of recommendation system using big data with a clear explanation in representing the operation of mapreduce. We demonstrate the various stage of recommendation namely data collection rating, types of filtering. Analysis Scenario based drug recommender system, it consists of three components namely drug storage, cloud server, and recommender server. The system is evaluating with specific parameters like F-score, Precision, and recall. Finally, we describe the challenge of recommendation systems like data sparsity, cold start, sentimental analysis and No surprise.

1 Introduction

Currently, we move towards a new stage of information development, one it has been performed in Organization with a significant amount of real-time or near real-time data to enhance effects of critical results in a business environment. Nowadays, most of the information is improved faster than its ability in processing it. Every one has its uniqueness, but same time there are having more criticized relevant to each other, it demonstrates the same behavior, performing with the same user and liking same items. But it is not a requisite lousy idea. We required determining it advantages from collective intelligence that can combine with lots of application utilized on a daily basis. The user needed preferring the trust or opinion to assist them in identifying the item they preferred.

M. Sandeep Kumar (✉) · J. Prabhu
School of Information Technology and Engineering,
VIT University, Vellore 632014, TamilNadu, India
e-mail: sandeepkumarm322@gmail.com

J. Prabhu
e-mail: j.prabhu@vit.ac.in

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_44

2 Related Work

Harshawardhan et al. [1] represent the idea of big data using three V's namely volume, velocity, and variety. Heterogeneity, lack of structure, error handling, privacy and visualization are some of the technical challenges in big data.

Patel et al. [2] describe large dataset recommendation system using Hadoop framework. Rating, review, opinion, feedback are some data items available on the web. Mahout platform has used for implementing this recommendation system using Movielens datasets it offers both review and rating.

Adomavicius and Tuzhilin [3] represent various limitations, methods, and approach in recommender system and possible criteria to enhance the recommender system with wide range of application. We required knowing that enhancement of both user and item, combined with context information in recommender stage.

Zhao and Shang [4] represents user-based CF algorithm based on a cloud platform that performs to solve scalability issues by Hadoop. The process performed using MapReduce. It has a component like a mapper and reduces to partition. The advantage of mapreduce that it complete task at the same time with linear speed up.

Patil et al. [5] represents product recommendation filter system for usage of big data in Hadoop and spark. Spark perform with a large dataset, and it runs more than 100 times quicker than Hadoop MapReduce in memory, so requires to realize that this system of spark offers more scalability and real-time RS. This method also compared with the different algorithm in providing more efficient, accurate in rating estimation.

Isinkaye et al. [6] describe recommender system with different characteristics and predicting technique used in the order. It represents various stages of recommender process namely data collection-Explicit, implicit and hybrid feedback, learning stage, and Predication stages.

Nowadays diabetes mellitus is considered as common diseases. Based on a present survey of world health organization predicates number of diabetic improved more than 56% in Asia from 2010 to 2025, similar to number of drug medicine that doctor can use and also develop as pharmaceutical drug development. Chen et al. [7] introduced Diabetes medication RS depending on the field of ontology; it mainly applies the knowledge base offer to hospital specialist in the hospital. This system analysis knowledge of drug, attributes, type of consequence and ontology knowledge regarding patient symptoms.

Wang et al. [8] represent Novel RS based on drug repository strategy. The system will perform drug indication and its consequence in one combined operation. This approach offers a more additive method to medical genetics-based drug storage it will decrease the process of false + in the repository. Thus show a list of new person indication with various levels of confidence. Several further drug indication and consequence of side effect are documented with the more likely level of confidence.

3 Big Data Analysis and MapReduce

Quick improvement production of information from a different aspect, there also required changing a vast amount of unstructured data into need information for social usage. Whole information could be merely performable and make utilize to determine semantic interoperability in electronic activity. The term big data refers to large datasets that could be analyzed and perform with the consistent pattern, trends, which are related to human behavior and interaction. It can take improve decision at the right time.

Most researchers frequently refer big data as three V's namely Volume, variety, and velocity [1]. Volume refers to the extreme amount of data to be determined, and analysis, Variety consist of both structured and unstructured data types and velocity refers to the speed of the data to be record and review based on action. While gathering knowledge of big data, we will not avoid the concept of Hadoop and mapreduce [9].

Mapreduce is a software system that is simple to write an application and processing a significant amount of data in a symmetric way on large clusters of hardware commodity with fault tolerance. The system will classify its input sets to chunks that process in map task entirely in a parallel direction and the similar method will classify output of map as input to reduce a task. Both functions will be stored in file system namely input and output. The mapreduces system performed and utilizes for developers to split the query into each layer in a dataset, into a chunk that determines those step using pattern divide hosts [10]. Mapreduce model mainly based on two functions, map() and reduce() [11]. Mapreduce libraries consist of an enormous programming language with various levels of optimization. Google technology provides name called mapreduce.

@@**MapReduce component** (Ref-Fig. 1)

Name Node—name node based on client and HDFS it placed information between the file systems and add, delete and manipulation depends upon updating of data.

Data Node—Based on two primary functions namely Job Tracker and Task tracker. It comprises set of data in HDFS and determines as a platform for being jobs, and other data will resort to local data among HDFS.

Job Tracker—Job tracker will allocate jobs and observes the assign tasks to task tracker.

Task Tracker—Task tracker will observe his task and generate a report to the job tracker.

Mapreduce process (Ref-Fig. 1)

Maps reduce issues based on instability that contribute to improving distributed systems. The map will divide into multiple tasks namely mapper and reducer. In mapreduce consist of master and slaves. The master will save as "masters" configuration files, and slave saves as "Slaves and comprehends with each other.

Mapper

Mapper maps based on two keys such as input and intermediate key-value pairs. Mapper primarily depends on sorting process with more accuracy.

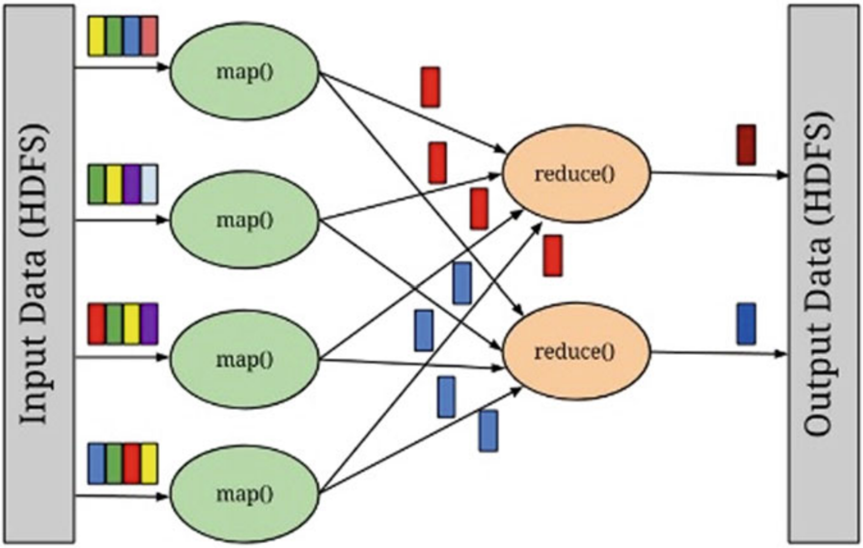


Fig. 1 Process of mapreduce

Reducer

The reducer will decrease it cut of value that are assigned as key to a small set of values. Shuffle, sort, and secondary sort are the three main stages of reducer.

4 Recommendation System

Currently, most of the real-time recommendation system perform using Big data scenario. Due to web offering a massive amount of information at present is more difficult for the user to identify required/similar information. Mainly big data related to data that are utilized for recommender system and filtering techniques. Moreover, big data is needed to allocate its standard benchmarking to it [12]. Frequently benchmarking required knowledge in deciding at the end to describe why benchmark needs and frequently not scalable with benchmarking products. Mostly large data need to lie with the support of decision makers. Recommendations limit with the hard decision for some recommendations. The system will solve these issues, and it has many impacts to reproduce in many ways.

Big data permit us in making recommendations with large data that we did not view before [2, 3, 13]. Example—Google search algorithm and Amazon-based reader action from a different reader. Both systems perform with an algorithm that learns from previous data. Recommender system determines with benchmarking due to it does not require an analyst at the final stage. It conveys with few data points from massive data. Sensibly most of the organization are performing product recommendation in one way or another.

The primary goal of a recommendation engine to estimate user opinion based on historical data are implicit and explicit feedback. They will modify application at present with most similar data to users. In consequence of the benefit of similar recommendation that has been improving in likelihood for a customer to place action, when we required growing with standardizing recommendation engine with contextual information. This information may assist to counterpart the recommendation, delivery way from historical process. An especially real-time recommendation system is providing more values among entire Organization part, it comparatively needs sophisticated environment. Amazon, Google, Netflix, Medical care are some organization frequent analysis and perform with real-time recommendation system.

Use of recommendation based on personal opinion of the users that are more effective in finding interest on an individual item from selected choices. Most of the researchers are trying the recommendation system with collaborative filtering with various algorithms, tools, and methods. Each algorithm and tools will provide different result while performing with CF.

5 Performance of Recommendation System

Most of the recommendation system exploits in well-defined and logical points that consist of data collection, rating, and filtering [13].

5.1 Ratings

Rating is most important aspects of recommendation system that user analysis is regarding the excellent product. When user compactable with a specific product, the user will act like adding to shop card, purchasing, providing a rating. The recommendation system may allocate implicit rating depends on user actions.

5.2 Filtering

The main goal of filtering concept is to reduce products based on rating and different user data. Recommendation system consists of three types of filtering namely, Collaborative, user-based and hybrid filtering approach. In collaborative, we determine user preference is performed, and recommendation will be provided. In user-based filtering, user's browsing history regarding likes, purchase, and rating are considered as record ahead in offering recommendations. Most of the companies preferred to use a hybrid approach; it has a combination of both collaborative and content-based filtering. Currently, Netflix makes utilization of a hybrid approach for better performance. Collaborative filtering techniques construct a model from customer's previous action as well as relevant decision produce by other customers, and then utilize that model to evaluate items that the user may be referred to it [2, 14].

6 Evaluation of Recommender System

Evaluation of RS based on criteria that have been adopting from information retrieval and signal detection theory. Some of the typical metrics like Precision, Recall, F-measure, ROC curve and RSME [15–17].

6.1 Precision

The percentage of applicable items are specifically preferred out of all recommended items.

6.2 Recall

The computation of an item is preferred from all applicable items, it also considers sensitivity.

6.3 F-Measure

F-measure considers as quality mean the difference between precision and recall. It integrates both precision and recall measure into a single metric.

6.4 ROC-Curve

The ROC (Receiver operating characteristic) is drawn to visualize to modify the positive value against negative value based on sensitivity threshold. Most of the sensitivity value performs form output of algorithms like 1 (preferred) or 0 (Not preferred). The ROC Curve assists in achieving the optimal threshold and comparing various algorithm against each other independently for picking the limit.

6.5 RSME

The root-square mean error used to analysis in comparing estimation against real data. By evaluating the square error for entire items and considering the root of the mean of the square error, we acquired values from sanction high deviation and comparatively providing less difference. When weight score is increased, the estimation may vary higher among real benefits, and less evaluation offers more accuracy

7 Case Study—(Drug Recommendation System—Medicines)

Due to the quick improvement of e-commerce, most of the user preferring to buy medicine online for the sake of comfort. Moveover user does not realize issue while purchasing drug without the knowledge and proper guideless. Major problem while purchasing medicines for online site namely Validity and reliability. Validity represents based on risk factor in buying the noneffective medicine for an e-Commerce site. Basically, Reliability, is not having any proper knowledge about medicines. It may lead to the serious trouble to the user using medicines. Drug recommendation system will prefer top related drugs based on symptoms to users. The first drug required to categories in various groups (Cluster) namely function description information and quality of experience (QoE). The system will determine based on collaborative filtering but it has a specific issue like performance cost, cold start, and data sparsity. Drug recommendation system has a combination of users, symptoms, and medicine [18].

7.1 Drug Recommendation System

7.1.1 Drug Storage

Entire data of different drugs has been gathered from various online sources and social networks. Cleaning data, clustering based on specific features, indications, characteristics are steps followed in preprocessing (Fig. 2).

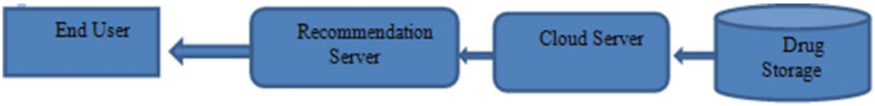


Fig. 2 Drug recommendation system

7.1.2 Cloud Server

Mainly used to determine details of drugs and present rating of the customer, we required demonstrating offline recommendation model based on tensor, meanwhile new item or drug rating updates will be placed in the model (Fig. 2).

7.1.3 Recommendation Server

Medicine name, symptoms are some keyword need to provide by the end user as input. Recommendation server will identify top N individual drugs in storage by this model (Fig. 2).

7.2 Evaluation in Drug RS

Data attributes of drugs namely name, description, username, age, and user rating. Evaluation of drug recommendation primarily depends on precision, F-score, and Recall [18].

Equation (1) refers to calculating accuracy, precision represented as

$$\text{Precision} = \sum_{i=1}^{M} \frac{\text{hit Medicine } i}{M} \tag{1}$$

M is indicated number of user's, hit medicine i = rug related with labels from recommendation list (tuple number)

Equation (2) refers to

$$\text{Recall rate} = \sum_{i=1}^{M} \frac{\text{hit Medicine } i}{\frac{R_i}{M}} \tag{2}$$

M is indicated number of user's,

hit medicine i = drug related with labels from recommendation list (triple number)

R_i Indicates the real triple number of the individual user.

Equation (3) refers to F1 score method used to evaluate weight of recommendation results

$$\text{F1} = \frac{2 * \text{precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

8 Challenges of Recommendation System

8.1 Sparsity

Most of the user will not rate the items and resulted in rating matrix turns into very sparse. Because data sparsity issues may arise that refuse the modification of finding a set of users with relevant ratings [19–21]. This will frequently be a drawback of the CF approach.

8.2 Cold Start

Cold start represents in gathering new item and users to RS. A new item will not able to prefer at starting stage, but it introduced to CF with no ratings. For example, Movielens will not favor for the new item until it gets some initial rating. Issues of new-user are hard to deal due it not having any possibility of identifying relevant users [22].

8.3 No Surprise

When we have sufficient data and recommend engines, perform critically there may be no changes [23]. Evaluation of recommender system and its algorithm is underlying with various issues for many reasons. (1) The different algorithm may be good or better depending on its multiple datasets. (2) Variation in evaluating different datasets. Retail and media industries are the two getting more advantages in using recommendation engines due to it has more data with a long tail and can invoke from cold start problem.

8.4 Sentiment Analysis

Mostly sentimental analysis related to opinion mining, it refers to opinion, preference, sentiments, and emotions expressed in the form of text. The main aim to determine in "translating different human action into hard data." Most of the researchers commonly preferred in scraping blogs and other related social media materials. Sentimental analysis helps to identify the quantifying whether, how active users will be happy or unhappy, pessimistic versus optimistic, like versus dislike, affirms versus decline among them.

Polarity and intensity are two true intent statements for entire fundamental challenges. Many obstacles can be prevented from this use of slang local dialect and irony to the deficiency of any key-term [24]. For improving sentiment analysis algorithm, required to deal with technical and measurement challenges. But its process with non-trivial at the stages is analysis in conceptual and classification. Finally, it lies in deciding for instance whether the presence of key-term or frequency. The input may be critical for the human analyst. Both central and generic of entire concept practices depends on classification. They will not be any enhancement without classification in reasoning, language, data analysis, and social science.

9 Conclusion

For business perspective, recommender system and big data play an essential role in many organizations. System permit user to pick items based on its interest, it plays like affirm a role in improving sales in organization. Currently, a recommender system is placed with final dataset available on sites, so we need technique or approach for enhancing the scalability of the recommendation system. In this paper, we explicitly explain the concepts of the recommendation system, big data, and mapreduce, etc. Demonstrate drug-based recommender system using some scenario and also represent parameter that is used to evaluate drug recommender system. Finally, describes the challenges of a recommender system.

References

1. Bhosale, H.S., Gadekar, D.P.: A review paper on big data and Hadoop. Int. J. Sci. Res. Publ. **4**(10), 1–7 (2014)
2. Verma, J.P., Patel, B., Patel, A.: Big data analysis: recommendation system with Hadoop framework. In: 2015 IEEE International Conference on Computational Intelligence and Communication Technology (CICT), pp. 92–97. IEEE, Feb 2015
3. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans. Knowl. Data Eng. **17**(6), 734–749 (2005)
4. Zhao, Z.D., Shang, M.S.: User-based collaborative-filtering recommendation algorithms on Hadoop. In: Third International Conference on Knowledge Discovery and Data Mining, 2010. WKDD'10, pp. 478–481. IEEE, Jan 2010
5. Patil, S.N., Deshpande, S.M., Potgantwar, A.D.: Product recommendation using multiple filtering mechanisms on Apache spark (2017)
6. Isinkaye, F.O., Folajimi, Y.O., Ojokoh, B.A.: Recommendation systems: principles, methods, and evaluation. Egypt. Inf. J. **16**(3), 261–273 (2015)
7. Chen, R.C., Huang, Y.H., Bau, C.T., Chen, S.M.: A recommendation system based on domain ontology and SWRL for anti-diabetic drugs selection. Expert Syst. Appl. **39**(4), 3995–4006 (2012)
8. Wang, H., Gu, Q., Wei, J., Cao, Z., Liu, Q.: Mining drug-disease relationships as a complement to medical genetics-based drug repositioning: where a recommendation system meets genome-wide association studies. Clin. Pharmacol. Ther. **97**(5), 451–454 (2015)
9. Jamiy, F.E., Daif, A., Azouazi, M., Marzak, A.: The potential and challenges of Big data-Recommendation systems next level application. arXiv preprint arXiv: 1501.03424 (2015)

10. Wali, M.N., Sree Prasanna, K., Surabhi, L.: An optimistic analysis of big data by using HDFS
11. Dhavapriya, M., Yasodha, N.: Big data analytics: challenges and solutions using Hadoop, map reduce and big table. Int. J. Comput. Sci. Trends Technol. (IJCST) **4**(1) (2016)
12. Bollier, D., Firestone, C.M.: The promise and peril of big data, p. 56. Aspen Institute, Communications, and Society Program, Washington, DC (2010)
13. Helbing, D., Frey, B.S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., Zwitter, A.: Will democracy survive big data and artificial intelligence? Scientific American. 25 Feb 2017
14. Azar, A.T., Vaidyanathan, S. (eds.): Advances in Chaos Theory and Intelligent Control, vol. 337. Springer (2016)
15. Wiesner, M., Pfeifer, D.: Health recommender systems: concepts, requirements, technical basics, and challenges. Int. J. Environ. Res. Public Health **11**(3), 2580–2607 (2014)
16. Yang, S., Zhou, P., Duan, K., Hossain, M.S., Alhamid, M.F.: emHealth: towards emotion health through depression prediction and intelligent health recommender system. Mob. Netw. Appl. 1–11 (2017)
17. Holzinger, A.: Machine learning for health informatics. In: Machine Learning for Health Informatics, pp. 1–24. Springer International Publishing (2016)
18. Aznoli, F., Navimipour, N.J.: Cloud services recommendation: reviewing the recent advances and suggesting the future research directions. J. Netw. Comput. Appl. **77**, 73–86 (2017)
19. Sharma, S.K., Suman, U.: An efficient semantic clustering of URLs for web page recommendation. Int. J. Data Anal. Tech. Strat. **5**(4), 339–358 (2013)
20. Ma, H., Yang, H., Lyu, M.R., King, I.: Sorec: social recommendation using probabilistic matrix factorization. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, pp. 931–940. ACM, Oct 2008
21. Melville, P., Mooney, R.J., Nagarajan, R.: Content-boosted collaborative filtering for improved recommendations. In: Aaai/iaai, pp. 187–192, July 2002
22. Schein, A.I., Popescul, A., Ungar, L.H., Pennock, D.M.: Methods and metrics for cold-start recommendations. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 253–260. ACM, Aug 2002
23. Sagiroglu, S., Sinanc, D.: Big data: a review. In: 2013 International Conference on Collaboration Technologies and Systems (CTS), pp. 42–47. IEEE, May 2013
24. Das, A.S., Datar, M., Garg, A., Rajaram, S.: Google news personalization: scalable online collaborative filtering. In: Proceedings of the 16th International Conference on World Wide Web, pp. 271–280. ACM, May 2007

Design of EHR in Cloud with Security

R. Prathap, R. Mohanasundaram and P. Ashok Kumar

Abstract The adaptation of data and resource sharing in health care is more popular and maintaining of Electronic Health Records (EHR) is efficient to doctors and patients. Nowadays storing of health records in cloud environment is more popular and low cost with scalable of data to access. But the security of those data assurance will be less in current technology. Health care started using a storing of data in cloud with minimal level of security. We propose method to provide data security and access control to the constraints patients or doctors to access the data using encryption of unique identification. We introduce a multi-user encryption with authorization.

1 Introduction

In healthcare system maintaining of Patients records, Doctors records, Diseases records are through the computer machine. Even though they are facing some critical issues related with maintaining records, health records are critical to the grand version of healthcare digitalization [1]. To improve the handling of patient in efficient way and reducing healthcare delivery cost [1] in the system, we introduce integration of healthcare records into the cloud with security. Cloud is having services and delivery models to offer on-demand self-service based on the pay-per-use go in the services. Electronic Health Records (EHR) is a new level of technology to maintaining healthcare records in cloud environment with high quality, affordable cost and interoperability of EHR in cloud environment.

In this paper, we proposed the design a cloud for EHR system and access control mechanism [1]. In healthcare organization uses minimum security for patient's login.

R. Prathap (✉) · R. Mohanasundaram · P. Ashok Kumar
VIT, Katpadi, Vellore 632014, Tamil Nadu, India
e-mail: prathap.r@vit.ac.in

R. Mohanasundaram
e-mail: mohanasundaramr@vit.ac.in

P. Ashok Kumar
e-mail: ashokkumar.p@vit.ac.in

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_45

Many algorithms are introduced for encrypting and decrypting process of logging the computer system. Virtualize information are easily hacked by the hackers or call it as third party. In order to avoid these problems, we provide high security enforced before deploying any cloud-based healthcare services [2, 3]. To provide more security to electronic health records and personally identifiable information with help of advanced security techniques to maintaining security to healthcare data. We perform the setup of private cloud environment with security of unique key generation to the end user. Dedicated cloud provides the information to every user to access the data in the cloud environment.

1.1 EHR Cloud in Healthcare

EHR cloud in health care is to develop a secure digital healthcare system to patients, doctors and so on. Second, to integrate global hospitals records in cloud environment. In daily life, huge amount healthcare data are generated. Each and every data must be important for care of patients. Cloud Computing is a virtualized storage of data in real time. It is one of the low-cost methods to storing and accessing the data. It is used to accessing and exchanging data between healthcare organizations [4]. The volume of healthcare data is high. So, in the cloud environment it is very hard to maintain security and privacy of patient's records, healthcare organization records in the virtualized platform [5, 6]. In order to avoid hacking of healthcare data in cloud [7], we introduce an efficient algorithm for storing, accessing, and exchanging healthcare records in the cloud environment.

It is one of the efficient techniques to avoid unwanted or third party accessing data in the cloud environment. An efficient algorithm is proposed to handling of personal health information of individuals or healthcare records in organization [8]. Thus, effective and advanced security algorithms introduce to avoid threats and hacking of EHR or PII in cloud environment.

Various algorithms and techniques are used in Cloud Environment to store and retrieve information saved in a cloud services with security Based on the analysis done using private cloud model, the cloud system can provide efficiency and improve reliability to take care of patients and resources needed users [9].

1.2 Existing Techniques Follows in EHR

The collection of medical records has increased aggressively. With electronic health records (EHR) [1], patient databases, and so on. Digital healthcare system in cloud evolving increased day to day in life [4]. Although the use of cloud in healthcare research remains in its process, it has to change the level of health care system and the integration of patient's records and hospitals information in the system.

In day to day, there is increasing generation of health record in the organization. So, they introduce an EHR (Electronic Health Record) to maintain information about the patients in the hospital [10]. Then nowadays many hospitals are move on to virtualized storing and accessing of healthcare data through network [5, 11]. Because, the technology cloud is a cost-effective model as pay use go in the environment.

In the cloud technology, many business deployment models are available to accessing the environment. Many hospitals moving into cloud environment because management of data was more difficult in the hospitals computers. In day to day growing of electronic health record and patient personal information is high.

Security and privacy of cloud environment is challenging task to the EHR (Electronic Health Record) or personal record [1, 11, 12]. Our efficient algorithm is used to provide the multiple number private key to the different level of users in the cloud environment.

1.3 Analysis of Cloud Service in Healthcare Field

The virtualization technology is good level of idea to health records maintaining health information is sharing in the global market.

Software Service:

The network or online software service of health and medical information service are software application performed by the hospitals. It can be maintained and updated through hospital admin. The main advantages are to reduce the buying on software license. Instead of that cloud will get only service cost from the organization [10].

Information Storage:

Information Storage service is provided by cloud service provider, here patient health records are integrated in the platform. Here management of hospitals information resources, patient healthcare records and it will effectively flow between the various organizations.

Benefits of cloud in medical fields:

Clinical Benefits:

The single clinical benefits can provide access to application, patient care can be improved by providing this service through the cloud faster and more efficiently. Since patients do not need to travel, waiting lists are more easily managed as more patients can have the some test in more locations with a larger availability of experts. These some experts can access patients data remotely and a demand through the network.

Business Benefits:

Cloud technologies provide huge benefits that can contribute to the welfare of a provider organization [5]. Healthcare provides are in the business of treating and costing for patients. The cloud offers providers the ability to access specific experts to manage and maintain their systems.

Reduce Costs:

It will enable primary healthcare provider to achieve higher efficiencies at lower cost.

Information Resources Sharing:

Healthcare work together to build medical information sharing space to form an infrastructure.

Hardware Reduce Operation and Maintenance Cost:

In this model user terminal configuration is no limit, the technical staffs does not have to upgrade by hardware. It reduces the work.

2 Overview of EHR Cloud with Security

Cloud Computing is a new virtual reality service and business model. The most significant feature is to provide service and business model to user. So, common user can access the computer resources in the Internet. Cloud Computing can provide business [8] model with software, to reduce hardware and software maintenance cost. Cloud Computing. As a recent developed technology, it will change the conventional computer resources application model, to change in business IT application. Recently, it will play a major role in medical field [13]. To maintaining patients records, operational resources details, hospitals maintenance, etc. Here, we create an application of cloud computer in medical field, especially to maintain EHR (Electronic Health Record) with security (Fig. 1). The major drawback in healthcare organization is PHI (protected Healthcare Information) [8] has theft from the computer.

In the private cloud service provider has different types of services, such as infrastructure, application, software and database services. In general user node interacts with Internet to requesting information about the patients or hospitals details in the cloud environment. They will access data using minimum level security of username and password to enter into the cloud. Our intention is to provide a security to end user giving unique key generation to the users [9].

3 Healthcare System in EHR with Key Generation

In the digitization of healthcare system, Public cloud and private cloud are interacting to accessing different kinds of services in the environment. Compare to public and private cloud, private is dedicated virtualized data center [14]. When hospitals information are merged with in private cloud, user is requesting information from

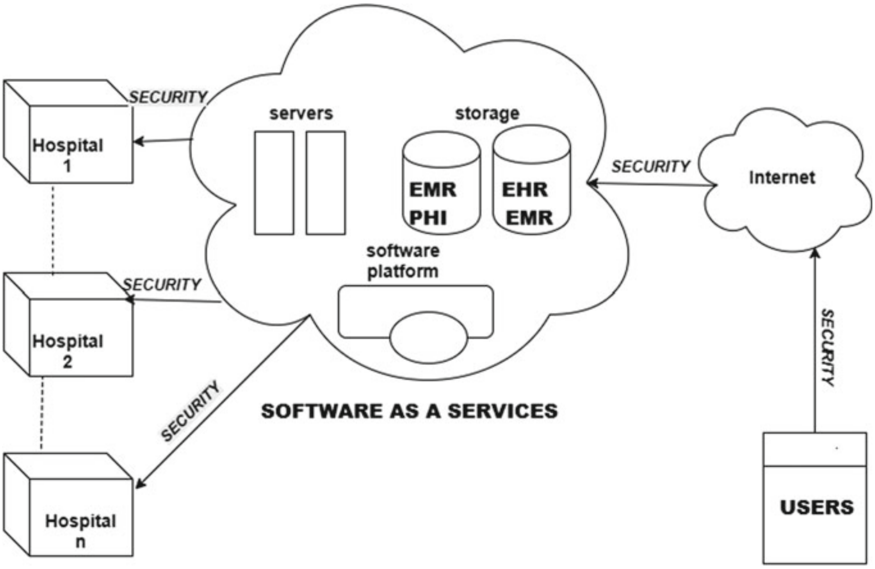


Fig. 1 Secure EHR in private cloud service

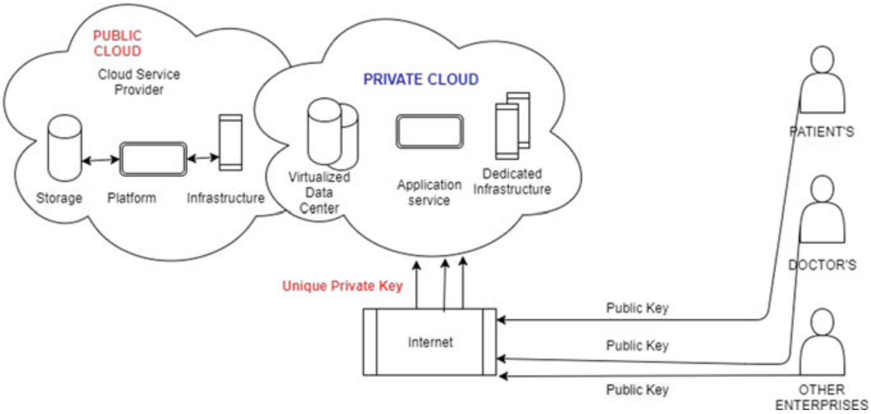


Fig. 2 New unique access code generation to end

the Internet, with base level security. By using advance algorithm, we will generate unique key to each nodes involved in the cloud. When the sender requests information from Internet and service provider replying to the request with minimum level security [15]. Mainly, purpose of private cloud is to use dedicated services to the patients, doctors and so on. In security point of view, they provide unique key to getting information in the cloud environment (Fig. 2).

Privacy and security is top issues in the healthcare data. Personal health information, accessing data in data centers patient privacy laws are concerned. The pos-

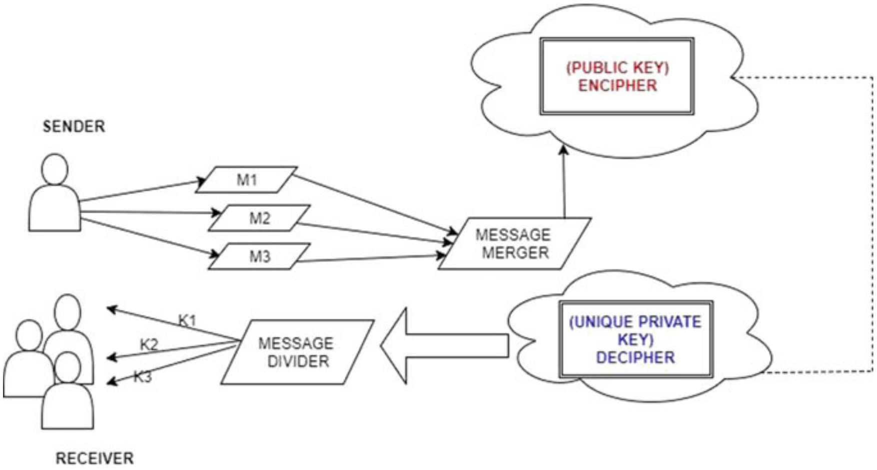


Fig. 3 New unique access code generation to end users

sibility of patient data could be lost, misused in wrong hands. Violation of patient confidentiality carries heavy fines, including significant cost of recovery and patient notification. The solution to provide private cloud.

In our approach, we make a private cloud environment with security of unique key generation to end users [1]. Dedicated data center provides information to every user to access the data in the cloud environment. Security Challenges in health and human service studies PHI violation have come from the theft of computer. Theft has been more for the computer and less for PHI (Fig. 3). In the healthcare organization for EHR, they are all using minimum security for patient login [13]. Many algorithms are also introduced for encrypting and decrypting process to the patient and organization record information [14]. We introduce multifarious private key algorithm for high security and privacy for individual's patients or any healthcare organizations. Based on the deployment model, we providing high end security to end user.

4 Conclusion

Cloud Computing network virtual service to everyone, such as e-mail, search engines, etc. By using search engine with simple keywords we get lots of information. Cloud computing in medical industry is to search on data or patient records, analysis of DNA structure, etc. To provide more security to EHR and PII and to provide security and privacy based on the deploying model using in the healthcare data. In cloud environment provide resources efficiently with security to the end users. Future work as to implement high-end security to advanced techniques based on the user requirement processes involved in the system.

References

1. Deshmukh, P.: Design of cloud security in the EHR for Indian healthcare services. J. King Saud Univ. Comput. Inf. Sci. **29**(3), 281–287 (2017)
2. Hu, H., Xu, J., Ren, C., Choi, B.: Processing private queries over untrusted data cloud through privacy homomorphism. In: 2011 IEEE 27th International Conference on Data Engineering (ICDE), pp. 601–612 (2011)
3. Sahafizadeh: Survey on access control models. In: 2010 2nd International Conference on Future Computer and Communication (ICFCC), vol. 1, pp. V1–1. IEEE (2010)
4. Huang, Q.Z.: Medical information integration based cloud computing. In: 2011 International Conference on Network Computing and Information Security
5. Geoghegan, S.: The latest on data sharing and secure cloud computing. Law. Order 24–26 (2012)
6. Li, H.-C.: Analysis on cloud-based security vulnerability assessment. In: 2010 IEEE 7th International Conference on e-Business Engineering (ICEBE), pp. 490–494. IEEE (2010)
7. Boyinbode, O.: CloudMR: a cloud based electronic medical records. Int. J. Hybrid Inf. Technol. **8**(4), 201–212 (2015)
8. Ved, V., Tyagi, V., Agarwal, A., Pandya, A.S.: Personal health record system and integration techniques with various electronic medical record systems. In: 2011 IEEE 13th International Symposium on High-Assurance System Engineering, pp. 91–94 (2011)
9. Xu, X., Li, L.: An artificial urban health care system and applications. IEEE Trans. Intell. Syst. **25**(3), 63–73 (2010)
10. Benaloh, J., Chase, M., Horvitz, E., Lauter, K.: Patient controlled encryption: ensuring privacy of electronic medical records. In: Proceedings of the 2009 ACM Workshop on Cloud Computing Security, pp. 103–114. ACM (2009)
11. Zhou: Security and privacy in cloud computing: a survey. In: 2010 Sixth International Conference on Semantics Knowledge and Grid (SKG), pp. 105–112. IEEE (2010)
12. Tong, Y., Sun, J., Chow, S.S.M., Li, P.: Cloud assisted mobile access of health data with privacy and auditability. IEEE J. Biomed. Health Inf. **18**(2), 419–429 (2014)
13. Wang, Y.-H.: The role of SAAS privacy and security compliance for continued SAAS use (2011)
14. Tang, P.C., Ash, J.S., Bates, D.W., Overhage, J.M., Sands, D.Z.: Personal health records: definitions, benefits and strategies for overcoming barriers to adoption. J. Am. Med. Inf. Assoc. **13**(2), 121–126 (2006)
15. Oza, N., Karppinen, K.: User experience and security in the cloud—an empirical study in the finnish cloud consortium. In: 2010 IEEE Second International Conference on Cloud Computing Technology and Science (CloudCom), pp. 621–628. IEEE (2010)

Comparative Analysis of a Systematic Coherent Encryption Scheme for Large-Scale Data Management Using Cryptographic Encryption Technique

A. Stephen Dass and J. Prabhu

Abstract In today's world of technology, Data has been playing an imperative role in many different technical areas. Data confidentiality, integrity and data security over Internet from different media and applications has greater challenges tasks. Data generation from multimedia and IoT data is the biggest Big Data in Internet now. When sensitive and confidential data are counterfeit by attacks this lead to serious countermeasures to security and privacy. Data Encryption is the mechanism to forestall this issue. At present, many encryption techniques are used for multimedia and IoT but when there is massive data are developed it incur more computational challenge. This paper design and propose new coherent encryption algorithm which holds issue of IoT and multimedia big data. The proposed system can cause a strong cryptographic effect without hold much memory and easy performance analysis. Handling huge data with the help GPU is included in the proposed system to enhance the data processing more efficiently. The proposed algorithm is compared with other symmetric cryptographic algorithms like AES, DES, 3-DES, RC6 and MARS based on architecture, flexibility, scalability, security level and also based on computational Running time, Throughput for both encryption and decryption process. Avalanche effect is also calculated for proposed system with 54.2%. The proposed framework holds to secure the multimedia against real-time attacks.

1 Introduction

Currently we are moving towards big data era. Data generation in day-to-day life is from the source comprises of social media data (like twitter, facebook, whatsapp, Instagram, etc.,) healthcare domain, business analysis and so on. For example in

A. Stephen Dass (✉) · J. Prabhu
School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, India
e-mail: mail2stephendass@gmail.com

J. Prabhu
e-mail: j.prabhu@vit.ac.in

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_46

2015 approximately 665 TB of data are manipulated and used in healthcare system [1]. When comparing with US library web of congress achieve is the huge massive size. On other side Data generation in social activities site rapidly increasing and also in query used in search engine for surfing in internet are huge by email transaction. Sharing images, videos, etc., data generated from navigation generated data by vehicles for tracking system using GPS. The Buzzword "Big Data" not only targeted on volume but also on their velocity and variety where we know about 3vs [2]. Moreover other behavioral characteristic such as value, veracity, validating, variability, venue, vocabulary, vagueness, verification, and visualization are often premediated and explained. For past two decades, data generation, data in vocabulary were not been safe by any other means of device leading in depreciation and loss of data. The organization were not able to store and manage all the repository data for long span of time, moreover organization was not able to handle these massive dataset. Unfortunately traditional data storage and processing are not able to handle these massive data in storing and managing the data due to its high cost expense. This lead to lack of flexibility, scalability to the system and also workout is mandate to enhance big data context. These IoT and multimedia data tends to produces 60% of internet traffic and 70% of mobile communication data [3]. This makes Big Data multimedia more complex. One of major factor affecting multimedia Big Data is security.

IoT (Internet of Things) is defined as structure or combination of objects such as sensor, receiver and transmitter which generates data. It has Internet work connection with Internet and other objects. IoT objects share their information with each other when they generate related data. Internet of things refers with things related to Internet. Many researcher and experts forecasted their views and ideas on IoT as by 2020; IoT Technology will have scheduled more of 50 billion objects as data generated things [4]. Nowadays many security systems have been developed and demonstrated for making communication medium safe in both IoT and multimedia data sharing file in Big Data. Many researchers are initiated using security concepts like cryptography method to secure data. They employ symmetric and asymmetric encryption techniques to solve security issue in Big Data. Among this public key encryption using single key transmission is most common security system. They are DES, 3-DES, AES, MARS, Blowfish and RC6.

2 Related Works

Putual et al. [5] proposed a novel key exchange method known as Dynamic prime number based on security verification (DPBSV). This Scheme especially developed for Big Data streaming which helps in improving the efficiency of verification production with reduced time interval. The system verifies for security for data stream manager. The authors manipulated the processing from stream processing Engine. DPBSV is design based on symmetric Key cryptography and random prime number generation. The future work of the author is do a comparative analytics study on proposed work with other symmetric key algorithm likely RC5, RC6, MARS. Furthermore Deepak wants enhance his research towards Internet of things.

Wang et al. [6] presented leakage volatile CP-ABE and KP-ABE scheme to improve the auxiliary model input. This scheme develops a strong improved extraction from Goldriech-Levis theorem to secure side channel attack. However this scheme secures the auxiliary input but fails in capturing from decryptor where leakage comes from PRNG weakness.

Sookhak et al. [7] provide an efficient remote data audit which is great advantage for auditor storing their value in cloud computing. This RDA is based on algebraic signature properties used in cloud storage security. They also put forth divide and conquer table which is efficiently supporting the data operation like append, insert, modify, and delete. This scheme can be applicable to large-scale massive data storage as future work. They intend to focus on auditing method for massive archive files in cloud storage system.

Xin [8] suggested a cross-combined hybrid scheme putting symmetric and asymmetric encryption to pledge with IoT-related data generation for high impact security. This scheme proposed with the help of Advanced Encryption Standard (AES) and Elliptic curve cryptography (ECC) because AES id defined in context of rapid, veracious Encryption technique for lengthy plaintext. Elliptic Curve Cryptography is used as key management and digital signature.

Prasetyo et al. [9] uses IoT data with symmetric key encryption technique. This scheme uses blowfish because of its fast and secure energy efficient for security IoT data. The proposed scheme examined the avalanche effect, Encryption time, and throughput to enhance the performance. The drawback of this system is that, usually use only IoT-related input files not the multimedia data as input files.

3 Proposed Model

In this module we instigate Systematic Coherent Encryption Scheme (SCES). It is composed of three parts 1. Hybrid key Schedule using MD5 2. Blowfish Encryption 3. Genetic Algorithm. The proposed system is explained in Fig. 1.

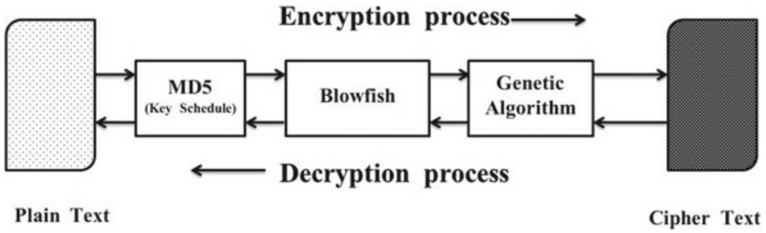


Fig. 1 Proposed system—MD5-blowfish encryption system explains the hybrid algorithm for processing the large-scale data

Blowfish is contrived by Brunce schree in the year 1993. It is one among the symmetric key block cipher. Blowfish with 64 bit block size plaintext and ciphered key with 32 bit to 448 bit is initiated as input file. They are 16 rounds included consist of key-dependent S-Boxes [10]. Blowfish is fast and highly efficient for securing data. This can be used by any type of user because it is unpatented, license-free. The main trait in blowfish is managing and key manipulating. This is achieved somewhere when each block is encrypting using generated key from data itself.

In cryptography, Block cipher is developed with symmetric structure like Feistal network. Feistal network is designed by Horse Feistal and Don Coppersmith in 1993. The original data is gives input file which is then partition into equal size block. Each and every part is categorized into plain text and key part of the input file. Here comes the role of Feistal network is to encrypt/decrypt the key scheme.

In genetic algorithm part, mutation and crossover are two ranges of combination process in genetic algorithm. In crossover, partial cipher text and ciphered key are swapped. In mutation, chosen bit is flipped in both ciphered text and key. The resultant is combined used as the Encryption file as whole cipher block. Furthermore crossover and mutation increases the system security against brute force attack.

3.1 Hybrid Key Schedule Using MD5

The main purpose of developing this key schedule algorithm of MD5-Blowfish is to have a high security of data without any data loss during encryption and decryption processing. This MD5-blowfish algorithm comprises of

1. Initial Key Encryption phase
2. MD5 Key Expansion Phase
3. Blowfish Data Encryption Phase.

Processing of key is followed in such a way that key is initial taken into key encryption Phase and then passed with Data Encryption phase.

3.1.1 Key Encryption Phase (KE$_n$P)

The user-defined key is encrypted using MD5 algorithm is fully loaded into Key Encryption Phase such that the initial key is encrypted by MD5 with key length of 64 bit size so it will be used to expansion module in Blowfish. MD5 is used as Key Encryption because it has no limitation on PlainText size to encrypt for key when compared with other algorithm like AES, DES, and MARS. Since they use multiple block size.

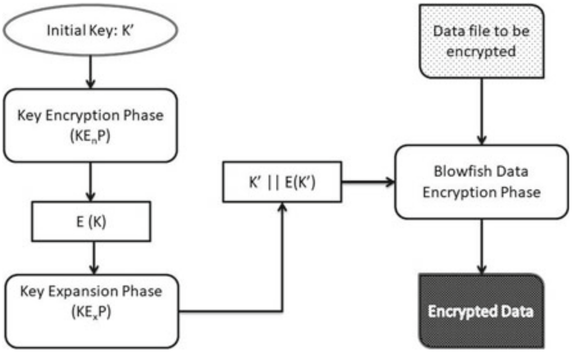

Initial Key: K'
Key Encryption Phase (KE$_n$P)
E (K)
Key Expansion Phase (KE$_x$P)
K' || E(K')
Data file to be encrypted
Blowfish Data Encryption Phase
Encrypted Data

Fig. 2 Proposed system: MD5 key expansion

3.1.2 Key Expansion Phase (KE$_x$P)

MD5 key Expansion phase initiates to develop encrypted key. This Encryption key is passed with Key Expansion Phase due to its complexity. This is explained in Fig. 2. Encrypted key E(K′) is tracked on with user key K′ which can have key size up to 448 bytes (K′ || E(K′)) [11]. This tracked key will be passed along the Blowfish Encryption Module. It makes the process hectic and more secured in the system develops where the key become much more complexity when compared to Blowfish key Expansion module.

3.2 Genetic Algorithm

Genetic Algorithm makes a major role in this proposed system. The CipherText and CipherKey are integrated and fused together positioned in order to perform crossover and mutation operation. Crossover is the process of operation in which two independent blocks are fused and trail their parameter in random manner. There are four types of crossover technique in general, according to the proposed system need and desire. Observation is needed in performing crossover and mutation operation the probability of mutation is derived in a bit as $\frac{1}{\text{length(block)}}$, where, length (block) is the length of CipherText or CipherKey.

The process of mutation and crossover highly denoted with number of generation gives the number of times mutation and crossover is performed. This type of high operations gives high secured integration and permutation when repeated process. One of the drawbacks found is while in motion or runtime, it struggles to process its actions.

4 Security Analysis and Implementation

This section deals with the system implementation and analysis of our proposed encryption algorithm. The experiment is implemented to prove that the proposed system provides equal balance on both security and performance

4.1 System Implementation

In this proposed system, we implement using PHP and HTML as simple frontend and put forth the standards with symmetric Encryption algorithms namely Data Encryption Standard (DES), 3-DES with 168 bit key, Advanced Encryption Standard (AES), MARS. Blowfish with 64 bit key length and Rivest Cipher (RC6). From these algorithms we calculate the running of the algorithm, throughput for encryption and decryption process and the avalanche effect. The result was calculated from many number of iteration and average output is calculated as resultant outcome.

All experiment is carried in high-end computing multimedia lab to process large multimedia data. Hardware and software specification includes IBM cell processor to perform this operation with the node of 32 GB memory and Intel Xeon Processor E5 2609 v4 with 1.7 GHz of 20 m Cache 85 W CPU [10].

GPU Processor:

To process in data with operation in Graphical Processing Unit (GPU), we use NVidia tesla M60 GPU. The specification of the GPU and is sorted below in block diagram explains in Fig. 3.

Data are collected from IoT multimedia data dataset with various heterogeneous formats. We have made many cross-estimation for the proposed system with different input file sizes with different interval time stamp such as [1 KB–10 MB], [10–100 MB], [100–500 MB], [500 MB–1 GB], [1–5 GB], [5–10 GB], [10–15 GB] and [15–20 GB].

4.2 Performance Analysis

Performance Analysis are estimated using the Running time of the algorithm and Throughput value of the algorithm of proposed system compared with AES, DES, 3-DES, RC6, and MARS. The outcome of the proposed system put forth and discussed in order to fetch performance metrics using Running time and Throughput of the handsome of symmetric key algorithm like MARS, RC6, AES, DES, 3-DES.

All algorithms used in this paper have both major advantage and minor weak key because of its Cipher and its components uses different criteria. Since all algorithms are not same in its structure, no. of encryption/decryption rounds, key size. For

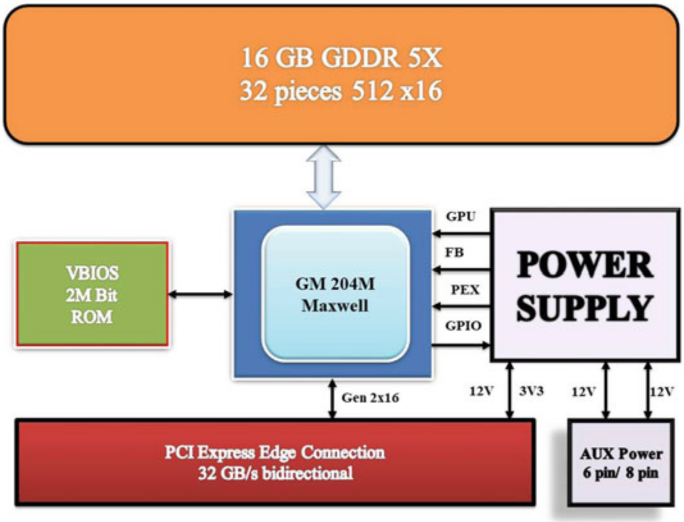


Fig. 3 NVIDIA GPU processor—Tesla M60 [12]

instance, we compare DES, 3-DES has same no. of rounds but 3-DES performs thrice the encryption of decryption and encryption process.

Based on Running Time

The proposed algorithm holds less running time when compared with other cryptographic algorithm. The algorithm runs in 6800 m/s for average time interval of [5–10 GB] of file size. The proposed system algorithm is well chosen to be secure and fast in practice. The combined Blowfish algorithm embedded with hardware and software add much more value in generating and processing large dataset to be secured. And also this proposed algorithm sustains against all known attacks performed by the intruders. Secondly AES hold least than proposed system tends to be secured. MARS has high security when compared with RC6 because of its key size (128 and 448 bits) whereas RC6 has key size of 128,192,256 and also has many number of rounds. In other aspect RC6 has more flexible than MARS because RC6 has less number of round than MARS, RC6 has 20 rounds whereas MARS has 32 rounds. And also RC6 uses less memory space when compared to MARS. MARS makes use of table lookup since like DES S-box with single table of 512 bit. Due to this it slows the software implementation in the system. DES has 56 bit key size whereas RC6 has 128 bit key size. The difference of key length determines the block to be encrypted and takes very less time. Finally we understand 3-Des has worse than MARS since 3-DES uses three time sequence with three different keys (K1, K2, and K3) and has 168 bits.

Running time in context with encryption, the time taken to encrypted definite file is directly proportional to the size of the original files. It explains and gives the average

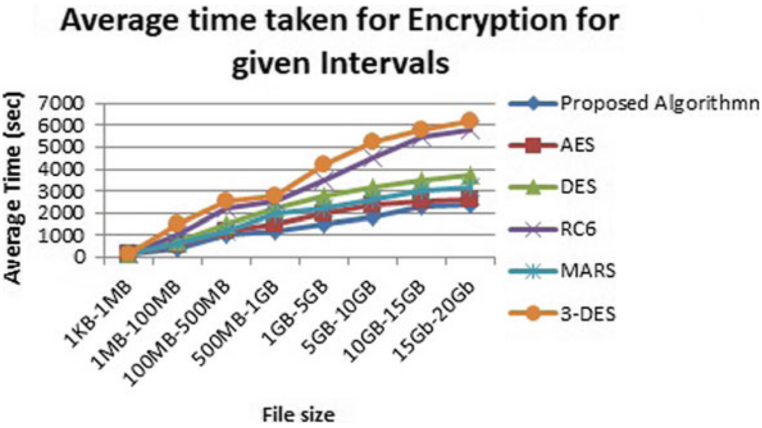


Fig. 4 Overall average time for encryption with given interval

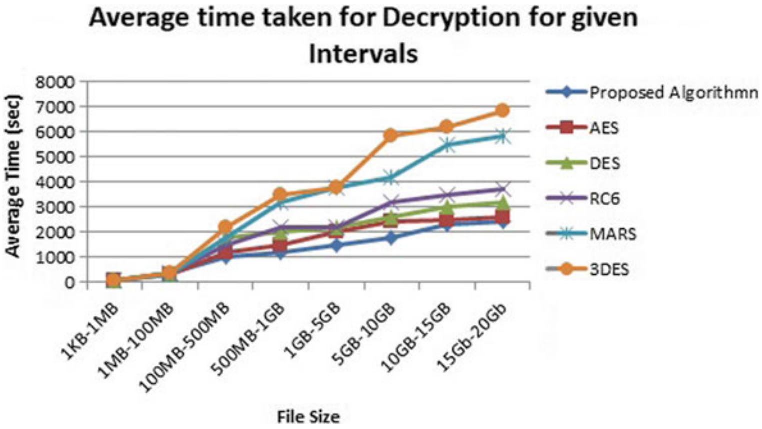


Fig. 5 Overall average time for decryption with given interval

time of the encryption done for input files in timestamp interval of ([1 KB–10 MB], [10–100 MB], [100–500 MB], [500–1 GB], [1–5 GB], [5–10 GB], [10–15 GB] and [15–20 GB]) (Fig. 4).

For decryption process Fig. 5 with input files explain the decryption process with the proposed algorithm. It decrypts with run time of 7100 m/s for average time interval of [5–10 GB] of file size.

Table 1 Performed processed algorithm with flipped on number of bits

Algorithms performed	No. of bits	Percentage (%)
3-DES	31	30.8
MARS	45	36.2
RC6	40	45.2
DES	37	48.5
AES	44	52.5
Proposed algorithm	65	54.2

4.3 Security Analysis

Security Analysis deals with evaluation of our algorithms based on security metrics. Using Avalanche effect we can conclude the security aspects issues. Avalanche Effect is estimated using the strength of the algorithm by how resist against the attacks by threats and real time attacks like brute force, etc. Avalanche Effect is defined as when an input is slightly changed. Consequently, there is a change in the output also. For instance flipped in single bit in encryption process changes the half of the output bit flip.

$$\text{Avalanche effect} = \frac{\text{No. of Flipping bits in the cipher text}}{\text{No. of bits in the cipher text}} \times 100$$

From Key Encryption algorithm context change in single bit plaintext make huge change in CipherText bit in order to find avalanche effect. Let us consider an example of two byte plaintext (1001 0010 1101 0011) is flipped with change in initial bit (0001 0010 1101 0011). After encrypting the original PlainText we get (1011 0110 1001 1110) and flipped CipherText is (0111 0110 1000 0110). From this we calculate the count of number of flipped in which is 8 divided with number of bits in CipherText which result in Avalanche Effect is 50%.

From Table 1 we draw a graphical representation of avalanche effect with algorithms like AES, DES, proposed algorithm, RC6, MARS, 3-DES. The outcome of these are taken into percentage such as proposed algorithm with 54.2% in highest and followed by AES with 52.5%, DES with 48.5%, RC6 with 45.2%, MARS with 36.2% and with least 3-Des with 30.8%. The proposed Blowfish + MD5 algorithm has cracked his high performance when compared with other algorithms.

From all the analysis and inference from Running time, Throughput and Avalanche effect, proposed algorithm holds the highest range in efficiency, fast, dynamic and scalable to large-scale data management (Fig. 6).

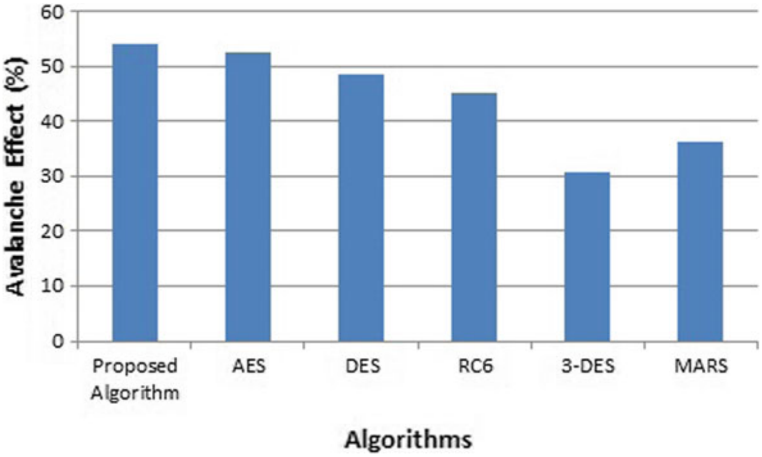


Fig. 6 Avalanche effect

5 Conclusion and Future Work

In this paper, we develop and implement a system with high reliability and dynamic GPU Encryption system for large multimedia IoT educational Big data resisting real-time attacks like DDoS, brutal force attacks and other tampering attacks. We made a keen observation on flipping the bits in resultant process and also in algorithm execution process with substitution and permutation of S-boxes in encryption and decryption process resulting in high avalanche effect. This is considered to be a novel encryption system combing two symmetric cryptographic algorithm used for large-scale data management to be highly secured. Future work proposed in this system is to develop a much more efficient algorithm to handle highly confidential government data dealing with IoT military and IoT Medical Multimedia Data.

References

1. Sejdić, E.: Medicine: adapt current tools for handling big data. Nature 507.7492, 306–306 (2014)
2. Laney, D.: 3D data management: controlling data volume, velocity and variety. Gartner (2001)
3. Atzori, L., Iera, A., Morabito, G.: The internet of things: a survey. Comput. Netw. **54**(15), 2787–2805 (2010)
4. Evans, D.: The internet of things: how the next evolution of the internet is changing everything. CISCO. Int. J. Internet **3**(2), 123–132 (2011)
5. Puthal, D., et al.: A dynamic prime number based efficient security mechanism for big sensing data streams. J. Comput. Syst. Sci. **83**(1), 22–42 (2017)
6. Wang, Z. et al.: ABE with improved auxiliary input for big data security. J. Comput. Syst. Sci. (2016)

7. Sookhak, M. et al.: Dynamic remote data auditing for securing big data storage in cloud computing. Inf. Sci. **380**, 101–116 (2017)
8. Xin, M.: A mixed encryption algorithm used in internet of things security transmission system. In: 2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC). IEEE (2015)
9. Prasetyo, K.N., Purwanto, Y., Darlis, D.: An implementation of data encryption for Internet of Things using blowfish algorithm on FPGA. In: 2014 2nd International Conference on Information and Communication Technology (ICoICT). IEEE (2014)
10. Suresh, Manju, Neema, M.: Hardware implementation of blowfish algorithm for the secure data transmission in internet of things. Procedia Technol. **25**, 248–255 (2016)
11. Wang, Z. et al.: Design and optimization of hybrid MD5-blowfish encryption on GPUs. In: Proceedings of 2011 International Conference on Parallel and Distributed Processing Techniques and Applications (2011)
12. https://computing.llnl.gov/tutorials/linux_clusters/gpu/

Attacks in Wireless Sensor Networks

Vyshnavi Nagireddy and Pritee Parwekar

Abstract In this paper our objective was to explore the routing attacks in each layer. Security is a major issue when we discuss about Wireless Sensor Networks (WSN). Wireless sensor networks (WSN) have a versatile environment. When the network components like sensors are deployed in unattended environment, they are then endangered to various attacks. Due to the innumerable applications, there is a massive scope for research in various fields of WSN. This paper enhances a survey on classification of security attacks and the defense mechanisms for each attack that are examined.

1 Introduction

Wireless sensor network is a group of autonomous sensors distributed over an area which communicate to record the changes in the environment which includes parameters like temperature, sound, pressure, humidity, speed, etc. A WSN is a vast network of resource-constrained sensor nodes with multiple preset functions, such as sensing and processing, to fulfill different application objectives. These sensor nodes are deployed in a structured or unstructured manner in a geographical area. A sensor node has a power unit, radio sensing unit, and a processing unit and components in a sensor network include a sensor, it is a device which detects a signal form a physical world or environment and responds accordingly and a node is a physical component of a network. The task of a node is to forward the information like receiving, sending, etc. All these components form a sensor node which can transfer and store information [1]. Through these sensors various applications can be categorized like Industrial automation, Traffic congestion, Waste management, Security and emergencies, Retail and logistics, Agriculture and health, etc. During any forest fires, the information is distributed among the sensor nodes and then communicated to the

V. Nagireddy (✉) · P. Parwekar
Anil Neerukonda Institute of Technology and Sciences, Bheemunipatnam, India
e-mail: nvyshnavi.6@gmail.com

P. Parwekar
e-mail: pritee.cse@anits.edu.in

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_47

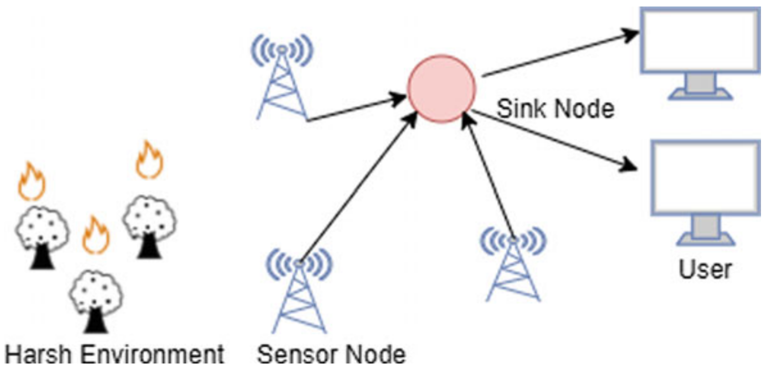


Fig. 1 Wireless sensor network architecture

users using a sink node as shown in Fig. 1. Apart from these physical components and applications, there exist various topologies namely Ring topology, Star topology, Bus topology, Mesh topology, Grid topology, Circular topology, etc. Deploying these sensors in a harsh environment leads to certain issues regarding the performance due to limitations like component failure, limited energy and transmission capacities, etc. [2]. Section 2 below explains about the classification of attacks. Section 3 talks about Critical analysis. Section 4 deals with challenges. Section 5 is Critical analysis and Sect. 6 gives the conclusion.

2 Classification of Attacks

Security is a major challenge to overcome when Wireless communication is considered. Wireless sensor networks are prone to attacks due to its mobility and energy constraints. Attacks can be categorized into active attack and passive attack. Passive attacks are passive in nature they cannot modify any information. Active attacks are menacing, apart from this, attacks can occur at the WSN layers like Physical layer, Link layer, Network layer, Transport layer, Application layer. Inter-process communication takes place in between the layers in OSI. In active attack is a chance of change in data and it harms the system. The change of data includes modifications, data traffic observation and information interruption [3]. This kind of attack is easy to detect than to prevent. These are similar to some DOS attacks in the OSI layers. In passive attack, it does not affect the network or the nodes, the information is altered. Passive attack is difficult to detect. Layer wise security attacks are discussed below in Fig. 2.

- **Jamming**: This occurs at physical layer. Jamming is a DOS attack. It is an attempt of making the users not possible to use network resources. WSN's are susceptible to jamming. Jamming degrades the network performance. Jamming is further clas-

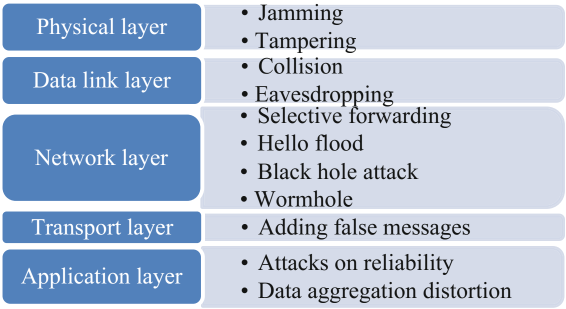


Fig. 2 Layer-wise attacks

sified into four categories they are Constant jammer, Reactive jammer, Deceptive jammer, Random jammer.

- **Tampering**: This occurs at physical layer. It is done to extract the confidential data. Attackers gain control over the path in a network and they gain access to it and steal the vital information like data through the nodes, Passwords and cryptographic keys [4]. Apart from getting all these information on the path, attackers can alter the information even.
- **Collision**: This is a DOS attack, in data link layer where a node induces a collision in some small part of a transmitted packet.
- **Eavesdropping**: It is an act that disturbs the communication between the source and destination. Credentials are lost in case of this eavesdropping attack. This attack occurs at data link layer.
- **Black hole/Sinkhole attack**: It is a kind of routing attack [5]. This attack is a menace to higher layer applications. A sinkhole attack creates a malicious node such that neighbor nodes gets attracted and think that the malicious node is the base station. The node is placed in a location where it can pull maximum traffic. This attack is difficult to terminate.
- **Sybil attacks**: In the Sybil attack, a malicious node behaves as if it holds a larger number of nodes. Single user pretends many fake/Sybil identities which are illegal here multiple identities are generated by a single node. Distributed systems are prone to this kind of attack. Effect can be direct or indirect [6]. Entities here communicate and identify each other using a communicating channel, that channel gets affected.
- **Denial of service**: Dos attack means making a resource unavailable for the user. Potential of sensor nodes is dissipated by the contender [7]. This attack can be expected in each of OSI layers. Jamming and tampering are the DOS attacks in physical layer attacks. DOS attacks in network layer are selective forwarding, sinkhole, Sybil, wormhole and hello flood attacks.
- **Wormholes**: In the wormhole attack, an attacker tunnels messages received in one part of the network over a low latency link and replays them in a different part of the network. It fake a path such that it appears as the shorter than the original route

within the network. This disrupts the regular routing phenomenon by placing a warm hole tunnel in between. It includes union of attacks like black hole attack, sink hole attack and eavesdropping [8].

- **Hello flood attack**: In this, the nodes with high radio transmission range and processing power sends hello packets to the neighboring nodes and creates a illusion that it is the nearest node. By this mechanism, attacker gain access over the information in the private network [9]. Attacker professes that messages are forthcoming from base station. Authentication can be a solution for this attack and it can be easily identified.
- **Selective forwarding**: It is a situation when certain nodes do not forward many of the messages they receive. The network often depends on forward mechanism that is carried out by broadcast messages which circulates throughout the network. Definite packets are dropped by a malicious node selectively. Dropping the received packets instead of forwarding can be referred as black hole at times [10]. Sequence numbers must be monitored regularly for a conjunction free network.

Other attacks irrespective of the layers are node tampering and packet dropping. A Packet is included with header and payload. Header's information is utilized by networking hardware to direct packets to its destination where the payload is extracted and used by application software [11]. Dropping of packets takes place over a network due to various reasons like network congestion and link congestion.

3 Critical Analysis

3.1 Flow Based Mitigation Model

This is one of the detection mechanism for sink hole attack. It includes various stages like Traffic log generation where there will be a log for packet forwarding and base station computes count of nodes which are represented in the table. The node here adds its own address before the packet reaches the destination. Next phase is Traffic pattern generation. Here the previous log form the first stage is considered. Memory limitations overcome here the old packets gets erased and constant time frames are kept in the traffic log. Next comes the Traffic log generation. It explores the condition and checks for presence of node if not a cycle message is sent and waits for the reply. Reply is passed on request through a long path [12]. When there is lengthy route, it is an indication of sinkhole attack.

3.2 Duty-Cycling Operation

It is a mechanism to detect the wormhole attack. It has two major segments like detection in time synchronization procedure and detection in synchronized communications [13]. The nodes broadcast messages with their respective timestamp and authority rating later comparison is done by considering the timestamp.

3.3 AOMDV Protocol

Ad hoc on-demand multipath distance vector routing protocol is used to locate the path linking source and the destination. This is addition to the AODV (ad hoc on-demand distance vector) protocol. MANET's use this AODV. Here routing table is maintained by each node and holds the information regarding the destination nodes. For communication between nodes, AOMDV checks the presence of route. It makes use of routing table to fulfill that task. This process is done to check whether there is a path between the two nodes. In the absence of route, request is made. On receiving this request destination sends a reply to source in the same path. This phenomenon builds various paths such that it becomes a rescue in any link failures [14].

3.4 Leach Protocol

It is a routing protocol in WSN. LEACH stands for Low energy adaptive clustering hierarchy which is a distributed clustering protocol. Leach is vulnerable to attacks and is used as a defense strategy for hello flood attack. Sensor nodes are clustered on the signal strength that is received. It has two phases called setup and steady-state. Clusters and cluster heads are formed in first phase. It is a repetitive process where sensors based on energy, strive to become the cluster heads. The status is updated to components in network by the cluster heads. In this phase finally nodes get sorted into cluster head and nodes are scheduled such that collision is avoided. Steady-State deals with the shifting of data from sensors to the base station in the network. Cluster head is responsible for data collection from sensors in the cluster. Data is then aggregated by cluster head to remove the data which is not reliable. These mechanisms are efficient as it extends the life time of network and the limitation is it cannot be applicable to larger networks [15].

3.5 Intrusion Detection System

The method proposed in this paper makes use of UDP (user datagram protocol). Data communication in between the nodes takes place using AODV protocol. Technique used here finds a solution for selective forwarding and hello flood attack which increases traffic among nodes in the network. This is accomplished using IDS and Key server. This IDS has command over the nodes and examines the actions done by the nodes under it. The work of Key server is to spread the public keys which are common and individual private keys to the nodes. Routing between the nodes is often a shortest path algorithm. When there exists any harm in the path, IDS needs to report the condition to key server. Key server then modifies the private keys for security [3]. Re-routing is performed by eliminating the effected node. Such node is not considered for further transmissions in the network.

3.6 Swarm Intelligence

Ant colony optimization attack detection (ACO-AD) algorithm which is a part of swam identifies the sink hole attack in sensor networks. Many approaches are prevailing to detect the sink hole attacks like swarm intelligence, Geostatistical models, Redundancy mechanisms, etc. Among this swarm is considered to be the best as it overcomes the pitfalls in neural networks and support vector machine architectures [5].

3.7 Sybil Attack Detection Technique

There are two phases in the Sybil attack detection technique. They are Location verification and Direction verification. Received signal strength table is maintained by detector in the network. It makes use of AODV protocol. In phase 1, identity is verified. Then the RSS identity is monitored and further computed, this is then compared to the stored identities and RSS values. If there is a mismatch then the node is identified with Sybil attack. In phase 2, query packet is sent to identities then we wait for the reply. If there is reply then the node is legitimate else there exists a Sybil attack [16].

3.8 Selective Forwarding Attack Detection Technique

It is a defensive mechanism for selective forwarding attack. It includes neighbor monitoring, attack detection, control packet collection, analysis and new path. In first phase packets are monitored whether they are dropped or sent or received by the base station. In the second phase, attack is detected in the base station by comparing the sequence number of the packets. This can be identified easily as the base station increments the count on drop of packets. In third stage the control packets are introduced as the base station requests for them [17]. Base station works on the control packets to determine the ID of defamatory nodes. Transmissions of control packets are verified by base station in the analysis phase. Moving on to New path

Table 1 Previous works on various attacks

S. No.	Authors	Title	Technique description
1	Devibala et al.	"Flow Based Mitigation Model for Sinkhole Attack in Wireless Sensor Networks using Time-Variant Snapshot" [12]	The method proposed here reduces the overhead and expands execution of the network
2	Takash et al.	"Poster: Detection of Wormhole Attack on Wireless Sensor Networks in Duty-Cycling Operation" [13]	Time difference is considered to detect the wormhole attack
3	Parmar et al.	"Detection and prevention of wormhole attack in wireless sensor network using AOMDV protocol" [14]	Round trip time and AOMDV protocol are the techniques used to prevent and detect Wormhole attack
4	Mayur et al.	"Security Enhancement on LEACH Protocol From HELLO Flood Attack in WSN Using LDK Scheme" [15]	By modifying LEACH protocol, a algorithm is proposed such that it works efficient for small networks to detect Hello flood attack
5	Anand et al.	"Localized DoS Attack Detection Architecture for Reliable Data Transmission Over Wireless Sensor Network" [3]	This paper presented the Intrusion detection system to resolve the Denial of service (DOS) attacks
6	Keerthana et al.	"A Study on Sinkhole Attack Detection using Swarm Intelligence Techniques for Wireless Sensor Networks" [5]	Swarm intelligence technique-Particle swarm optimization technique is a productive mechanism in detecting sink hole attacks
7	Khanderiya et al.	"A Novel Approach for Detection of Sybil Attack in Wireless Sensor Networks" [16]	AODV (Ad hoc on-demand distance vector routing) protocol is used in this paper for detecting Sybil attack. This method is energy effective as only one node is used under detection algorithm
8	Mathur et al.	"Defence against black hole and selective forwarding attacks for medical WSNs in the IoT" [17]	Mechanism proposed here is 96% efficient with improved accuracy against the selective forward attack

phase base station sends the information of defamatory nodes and sends a new path request which is free from defamatory nodes. The summary of attacks is given in Table 1.

4 Challenges

Majority of wireless sensor network frame work are deployed in a harsh environment. In such case, one can expect the performance degrade issues. The details discussed below are few of them:

- **Energy**: No wires are present in WSN, they are powered by a battery. In such case batteries are not reliable for a life time service in a node. They are to be recharged [10]. But in a harsh environment, it is difficult to inspect the nodes and replace the Batteries.
- **Multipath Fading**: This refers to the radio signal deformation [18]. As long as there exits at least one signal observation without experiencing a deep fade, the original transmit signal will be successfully decoded and recovered at the receiver, showing the reliability enhancement by exploiting multiple receive antennas.
- **Quality of service (QOS)**: Qos designed for WSN must reinforce to scalability. The node count should not affect the Qos. Minimum bandwidth must be providing such that one can have a better QOS.
- **Confidentiality**: Security is a primary issue here. One cannot continuously monitor the functioning of nodes or can trace the path.

5 Analysis

The survey here projects the overview of attacks in WSN and its countermeasures. Among all the security attacks in WSN, Wormhole attack is considered to be menacing. It is severe and destructive [19]. This attack totally destroys network topology. Often shortest path is preferred while transferring the data in between the nodes. On illusion created by the malicious nodes, they misguide with a short path. Data cycling and AOMDV protocols are efficient mechanisms revised in this paper for defense on wormhole attack. For the attacks in the physical layer, the defense mechanisms will be Tamper proofing and key management schemes [20]. Defense in data link layer is Encryption. In network layer, key management and secure routing are defense strategies. Application layer needs a unique pair wise key. These are the defenses available in the network.

6 Conclusion

In order to have a coherent data transfer, security of WSN must be considered. The work presented in the paper summarizes various security attacks and their countermeasures in WSN at each layer in the protocol stack. This leads to development in providing security measures to a Wireless sensor network which helps to overcome the challenges. The survey presented here optimistically aids the researchers to come up with effective defense strategies further for a secured network.

References

1. Bhatiya, A., et al. Detection and prevention of sink hole attack in aodv protocol for wireless sensor network (2017)
2. Shahzad, F., Pasha, M., Ahmad,A.: A survey of active attacks on wireless sensor networks and their countermeasures. arXiv preprint arXiv:1702.07136 (2017)
3. Anand, C., Gnanamurthy, R.K.: Localized DoS attack detection architecture for reliable data transmission over wireless sensor network. Wirel. Pers. Commun. **90**(2), 847–859 (2016)
4. Pooja, P., Chauhan, R.K.: Review on security attacks and countermeasures in wireless sensor networks. Int. J. Adv. Res. Comput. Sci. **8**(5) (2017)
5. Keerthana, G., Padmavathi, G.: A study on sinkhole attack detection using swarm intelligence techniques for wireless sensor networks. IRACST Int. J. Comput. Sci. Inf. Technol. Secur. (IJCSITS) **5**(5) (2015). ISSN: 2249–9555
6. Shabana, K., et al.: Security issues and attacks in wireless sensor networks. Int. J. Adv. Res. Comput. Sci. Electr. Eng. (IJARCSEE) **5**(7), 81 (2016)
7. Sandeep, Y., Hussain, M.A.: Implementing privacy homomorphism in data aggregation for wireless sensor networks. Indian J. Sci. Technol. **10**(4) (2017)
8. Mittal, H.: A survey: attacks on wireless networks. J. Netw. Commun. Emerg. Technol. (JNCET) **6**(5) (2016). www.jncet.org
9. Kurbah, R.P., Sharma, B.: Survey on issues in wireless sensor networks: attacks and countermeasures. Int. J. Comput. Sci. Inf. Secur. **14**(4), 262 (2016)
10. Grover, J., Sharma, S.: Security issues in wireless sensor network—a review. In: 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO). IEEE (2016)
11. Mesare, M.P.G., Sherekar, S.S., Thakare, V.M.: Prevention and detection of packet dropping and message tampering attack on MANET using EAMD (2017)
12. Devibala, K., et al.: Flow based mitigation model for sinkhole attack in wireless sensor networks using time-variant snapshot. Int. J. Adv. Comput. Electr. Eng. **2**(05), 14–21 (2017)
13. Minohara, T., Nishiyama, K.: Poster: detection of wormhole attack on wireless sensor networks in duty-cycling operation. EWSN (2016)
14. Amish, P., Vaghela, V.B.: Detection and prevention of wormhole attack in wireless sensor network using AOMDV protocol. Procedia Comput. Sci. **79**, 700–707 (2016)
15. Mayur, S., Ranjith, H.D.: Security enhancement on LEACH protocol from HELLO flood attack in WSN using LDK scheme. Int. J. Innovative Res. Sci. Eng. Technol. **4**(3) (2015)
16. Khanderiya, M., Panchal, M.: A novel approach for detection of sybil attack in wireless sensor networks (2016)
17. Mathur, A., Newe, T., Rao, M.: Defence against black hole and selective forwarding attacks for medical WSNs in the IoT. Sensors **16**(1), 118 (2016)
18. Zhu, J., Zou, Y., Zheng, B.: Physical-layer security and reliability challenges for industrial wireless sensor networks. IEEE Access **5**, 5313–5320 (2017)
19. Kurmi, J., Verma, R.S., Soni, S.: An efficient and reliable methodology for wormhole attack detection in wireless sensor network. Adv. Comput. Sci. Technol. **10**(5), 1129–1138 (2017)
20. Mulla, M.R.I., Patil, R.: Review of attacks on wireless sensor network and their classification and security. Imperial J. Interdisc. Res. **2**(7) (2016)

A Predictive Approach to Estimate Software Defects Density Using Weighted Artificial Neural Networks for the Given Software Metrics

T. Ravi Kumar, T. Srinivasa Rao and Sandhya Bathini

Abstract *Context* Applications of software have penetrated into every part of human life and have become significant in versatile areas. A case of negligence could cost a great loss of economy or even to life (Pentium FDIV bug "Statistical Analysis of Floating Point Flaw: Intel White Paper" (PDF) (2004) [1]; The Helminthiasis of the Internet [2]), and hence the reliability of the functions of an application is of extreme importance. A software error could be a deviation from the process flow, improper requirement definitions, or ambiguity in defining the constraints (IEEE Standard Glossary of Software Engineering Terminology (1990) [3]). To avoid the lateral effects, the applications should be evaluated beforehand for all defects in all phases of the software cycle. *Objective* This paper emphasizes on the prediction of software defect density using weighted neural networks and comparison with the fuzzy logic-based approach. *Method* Each design phase is divided into the neural network node. Software metrics are chosen for each stage and fed into the nodes with a weight associated to it. Neural network system are associated with input layers, SW-DDC layer (stage-wise defect density calculation layer) and C-DDC (consolidated defect density calculation layer). *Results* The accuracy of the predicted defects is very close the actual defects. Validation results show this method is very accurate than the current methods like fuzzy implementation. *Conclusion* Weighted artificial networks is used to predict the software defect density and is been verified on different real-time data sets ranging from very low project sizes to the very high project size.

T. Ravi Kumar (✉)
Department of CSE, AITAM Engineering College, Tekkali, India
e-mail: jhyma.gitam@gmail.com

T. Srinivasa Rao
Department of CSE, GIT, GITAM University, Visakhapatnam, India
e-mail: tsr.etl@gmail.com

S. Bathini
Appollidon Learning, Tampa, USA
e-mail: sandyvirgo25@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_48

1 Introduction

In the era of IT, the quality of the application should meet the client's requirements and should help in a healthy growth of the business. Ensuring the defect-less product deployable to the client is a great requirement for the developers. The standard's organizations such as IEEE/ISO too emphasize the need of reliability. Though it is hard to get the reliable application on the first go, certain predictions of the defects before deploying an application would help a lot. Experience on SDLC to get right data points would help here. The total number of defects is directly proportional to the size of the software [4]. Many metrics in each of the SDLC contribute to the identification of defects as early as possible [5]. Probabilistic approach and expert knowledge would help in the areas of uncertainty associated with software metrics [6]. Non-unified process in the application development drives the need of weighted artificial neural networks to determine the software density. The approach of the paper uses weighted inputs to neural network nodes and computes the defect density indicator which, in turn, provides the No. of defects for the given size of the project.

2 Software Metrics

Software metrics are the backbone of any model to decide the software reliability. In each phase of SDLC, there are so many metrics with effects the software quality. For this proposed model, we take software metrics are inputs to the proposed model. Initially, we define the weights of this metrics based on the expertise, as the learning of this model is evolved weights will be adjusted to make the predicted defects equals to the original defects.

Following is the thorough discussion of the input software metrics.

i. **Software size**:

Software size is the most important metric in deciding the software reliability. The size of the software is directly proportional to defects. This metric can be measured in lines of code (LOC). Usually, software sizes will be in KLOC. As the complexity of software increases, there is a high chance that software size gets increases.

Complexity α software size α possibility of error.

Stipulated development process and clean coding techniques can reduce the possibility of error even the complexity and software size increases.

ii. **Requirement software metrics**:

Requirement stage is the first and foremost important stage in the SDLC stable and clear requirement definition can highly impact the software quality. Among so many metrics of the requirement stage, the input layer is populated with Requirement stability, Fault density, and Review Information.

- Requirement stability (RS): Stable requirement is directly proportional to the software efficiency and reliability. Stable requirement gives a freehand for the designers and testers to concentrate on the freezed requirements. If the requirement change requests are more, i.e., unstable requirement disrupts the process in SDLC which in turn can cause the explosion of the software. Well-defined requirements can have adverse effect on the cost-quality reliability and schedule of the software [7].
- Fault density (FD): Fault density is inversely proportional to the software reliability. In requirement analysis phase, fault density can range from a low priority issue to the high priority issue which can impact the software in different ways along with increasing the probability of the defects. Continuous requirement analysis and early requirement description may reduce the fault density.
- Review Information (RI): Requirements review is directly proportional to the software reliability. Well reviewed requirement will eliminate any defects in the later stages. As found, the issue in the later stages of the SDLC will cost more in terms of cost and effort, it would be better to have a complete review of the requirements internally and externally with customers. Reviews should happen based on the software requirement specifications outlined by the prima organizations.

Requirement stage metrics impact on software reliability:

- Requirement stability α Software reliability.
- Fault density 1/α Software reliability.
- Review information α Software reliability.

Weights assignment for the requirement stage software metrics is $w_{\mathrm{RS}} > w_{\mathrm{FD}} > w_{\mathrm{RI}}$ where

- w_{Rs}: weight assigned for requirement stability.
- w_{FD}: weight assigned for fault density.
- w_{RI}: weight assigned for review information.

iii. **Design phase software metrics**:

Design phase follows requirement definition. Most of the cases coding start before even the requirement phase. Defect density of requirement phase is prime input to the design phase and also requirement phase defects affect the coding phase. Based on the literature survey mentioned above, software complexity and design review are considered as the software metrics.

- Software complexity (SC): Software complexity is inversely proportional to the software defects. Software program with larger number of decision points can make software more complex [8]. This is a good to have metric in the design phase since well-executed complex software attracts customers. It can provide a good indication for remaining software defects.
- Design Review (DR): Design review is to identify the defects or faults occurred in the design phase. Design review should make sure the design document outlines each and every detail of requirement specifications and match with the intent of

both. Whenever there is a requirement change, it should be promptly conveyed to design phase and get it reviewed. After each review based on the modifications if necessary, a re-review is suggested to make software quality assurance. Design review has a capability of eliminating defects from the later stages.

Design phase metrics impact on software reliability:

Software complexity 1/α software reliability design review α software reliability

Weights assignment for design phase metrics is $w_{\mathrm{DR}} > w_{\mathrm{RP}} > w_{\mathrm{SC}}$ where w_{DR} is weight assigned for design review$_{\mathrm{RP}}$ is weight assigned for requirement phase defect density and w_{SC} is weight assigned of software complexity.

iv. **Coding phase metrics**:

Coding phase metrics impact on software reliability:

Programmer capability α software reliability process maturity α Software reliability

Weights assignment for coding phase metrics is $w_{\mathrm{PC}} > w_{\mathrm{DP}} > w_{\mathrm{PM}} > w_{\mathrm{RP}}$ where w_{PC} is weights assigned for programmer capabilities, w_{DP} is weights assigned for design phase defects, w_{PM} is weights assigned for process maturity, and w_{RP} is weights assigned for requirement phase defects.

v. **Verification phase software metrics**:

Verification or testing is the critical phase in SDLC. This phase is useful to find the defects/bugs before software goes to the customers. In most of the scenarios, this phase contributes more towards the software reliability and quality. Staff experience and quality of tests are the prime metrics of this phase.

- Staff experience (SE): verification phase has a great impact on experienced staff. A technically sound and experienced staff contribution to the testing phase can directly impact the software quality. Globally organizations follow a pyramid pattern to balance the experience and knowledge domains.
- Quality of testing (QT): Tests are written to verify the software has to be designed well. As software testing is costly and time consuming, effective tests cases will assure a quality software time to market. Well-written tests will expose the software defects. Tests should cover the corner scenarios of the requirement. Good documentation of the issues or tests will ensure the software reliability.

Verification phase software metrics affect on software reliability:

Staff experiences α software reliability Quality of testing α software reliability

Weights assignments for testing phase is $w_{\mathrm{SE}} > w_{\mathrm{QT}} > w_{\mathrm{CP}} > w_{\mathrm{DP}} > w_{\mathrm{RP}}$ where w_{SE} is weight assigned for staff experience

w_{QT} is weight assigned for quality of tests, w_{CP} is weight assigned for coding phase defects, w_{DP} is weight assigned for design phase defects, and w_{RP} is weight assigned for requirement phase defects.

3 Proposed Model

The proposed model is the weighted artificial neural network. Model is divided into two major components, namely Defect Estimation System (DES) and Training module.

i. **Defect Estimation System**:

Defect Estimation System is a typical neural network with three different layers. Layers being Input layer, Hidden layer named as Stage-wise defect density calculator and Output layer which is consolidated defect density calculator. At the output, we will get the overall defects predicted based on normalization.

ii. **Input layer**:

Input layer is having nine nodes. Software metrics from the SDLC are fed into the input layer. Software metrics are chosen based on the literature survey. Each input is assigned with the corresponding weights. Weights being assigned to the inputs we can classify which metric is driving the software defects and evaluate necessary steps to ensure software quality. In future, if we need any other metrics to be considered it is easy to employ in the DES.

iii. **Output layer (C-DDC)**:

Inputs to the output layer are from the SW-DDC. As we had four nodes in the SW-DDC, we get four defect densities as the inputs. This layer has only one node to consolidate all defect densities and provide a cumulative defect density.

iv. **Normalization Process**:

To have inputs and outputs with the range, we follow a normalization process. This normalization is w.r.t to the software size, we normalize inputs with different levels based on the metric severity and impact on the software reliability. The following illustrations describe the normalization input software metrics. Training Module (TM):

Weighted artificial intelligence system is trained with around 30 real-time data sets taken from the promised [10] database. Training module assigns weights to achieve local and global minima of the error delta. Where error delta is been defined as the difference between the original defects and the predicted defects.

v. **Results**:

As shown in Fig. 1, weighted ANN-based system is predicting the No. of defects is very close to the actual defects compared to the most of the methods [12]. Results shown in Table 1 are simulated using Mat lab neural network toolkit. -Real-time data is been taken from the promised database [11]. In Fig. 2, we have compared our results with fuzzy-based software defects prediction model [9, 10]. As you can see in the table, most of the defects are predicted very accurately in the range of ±2%. ANN outputs after the normalization is rounded off to the nearest integer as there cannot be decimal defects. The data clearly depicts the relation between the software size and the number of associated defects.

Table 1 Inputs and output predicted outputs

RS	FD	HI	CC	DR	PC	EM	SE	QT	LOC(K)	Original defects	W-ANN Results[a]
0.16	0.87	0.83	0.5	0.75	0.85	0.93	0.87	0.93	6.2	383	365
0.63	0.15	0.39	0.44	0.19	0.095	0.51	0.033	0.35	9.8	589	592
0.75	0.49	0.82	0.93	0.23	0.62	0.15	0.65	0.42	2.4	172	177
0.89	0.73	0 94	0.17	0.89	0.84	0.77	0.92	0.88	3.6	236	239
0.35	0.034	0.074	0.57	0.56	0.39	0.37	0.27	0.13	7.8	417	425
0.2	0.36	0.72	0.48	0.67	0.19	0.87	0.56	0.34	12.8	640	634
0.41	0.54	0.86	0.44	0.98	0.34	0.67	0.238	0.864	30	728	727
0.65	0.42	0.23	0.15	0.23	0.62	0.35	0.65	0.84	34.7	751	758
0.92	0.88	0.89	0.77	0.89	0.84	0.2	0.27	0.39	42.5	821	825
0.27	0.13	0.56	0.37	0.56	0.39	0.37	0.56	0.19	51.6	975	974

[a] Rounded to nearest integer

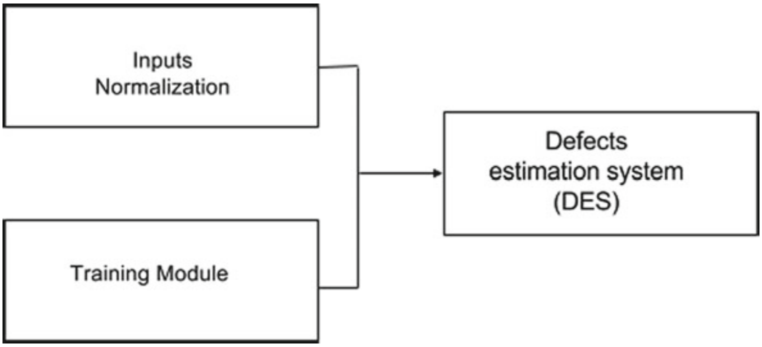


Fig. 1 Weighted ANN system to predict software defect density

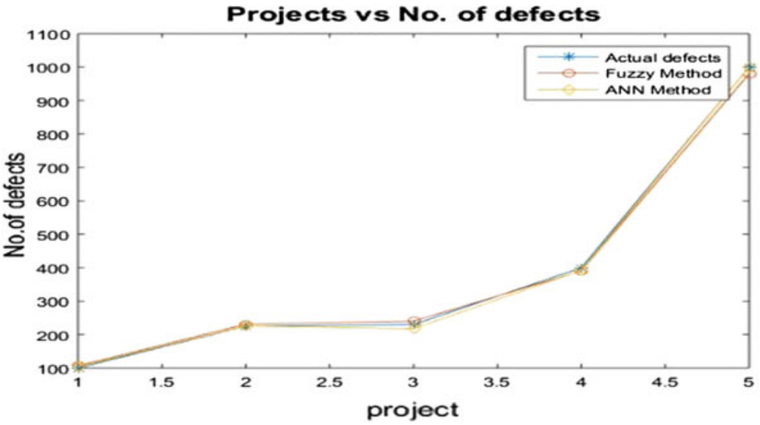


Fig. 2 Output function comparing the No. of predicted outputs

Model Validation:

Validation of prediction results is an evaluation metric for any proposed model [12].

i. Mean Magnitude of Relative Error (MMRE):

Absolute percentage of the absolute error is defined as MMRE. It is the measure to identify the spread of absolute error which is (estimate defects)/(Original defects).

$$\text{MMRE} = \frac{1}{m}\sum_{j=1}^{m}\frac{(x_i - \bar{x})}{x}$$

where x_i is the actual value and $\bar{x}$ is the estimated value of a variable of interest.

ii. Balanced Mean Magnitude of Relative Error (BMMRE):

Table 2 BMRE and MMRE values

Error rate	W-ANN
MMRE	0.015397
BMMRE	0.007258

MMRE is unbalanced and assesses over rates in excess of underrates. For this reason, a balanced mean magnitude of relative error measure is also considered which is as follows:

$$\text{BMMRE} = \frac{1}{m}\sum_{j=1}^{m}\frac{(x_i - \bar{x})}{\min(x_i\bar{x})}$$

For any proposed model, BMMRE and MMRE values show the model prediction accuracy. The lower these value higher the accuracy of the model. Table 2 captures weighted neural network-based software defect density registers the BMMRE and MMRE values.

4 Conclusion

In this paper, weighted artificial neural networks are proposed to identify the software defect in SDLC. Input software metrics to this model is chosen based on the literature survey. Weights assigned to each input which helps in tuning the system to get accurate No. of software defect. Tuning of this model is based on the training module. Training module is fed with real-time data from the promise repository. Input and output normalization is done based on the software size (KLOC). Defects predicted using this model is more close to the original defects and any other model. The validation measures BMMRE and MMRE shows very less values to indicate the prediction results accuracy. Results emphasize the use of Weighted artificial neural networks for the prediction of software defect density and ensuring software reliability.

References

1. Pentium FDIV bug "Statistical Analysis of Floating Point Flaw: Intel White Paper" (PDF). Intel. 9 July 2004, p. 9. Solution ID CS-013007. Retrieved 5 Apr 2016
2. The Helminthiasis of the Internet https://tools.ietf.org/html/rfc1135
3. IEEE Standard Glossary of Software Engineering Terminology. IEEE Std 610.12-1990,1,84 (1990)
4. IEEE Guide for the use of IEEE Standard Dictionary of Measures to Produce Reliable Software. IEEE, New York, IEEE Std. 982.2-1988 (1988)

5. IEEE Standard Glossary of Software Engineering Terminology. IEEE, New York, IEEE Std. 610.12–1990, pp. 1–84 (1990)
6. Kaner, C.: Software engineering metrics: what do they measure and how do we know? In: 10th International Software Metrics Symposium, vol. 6 (2004)
7. Kan, S.H.: Metrics and Models in Software Quality Engineering. Addison Wesley Publications (2002)
8. Diaz, M., Sligo, J.: How software process improvement helped Motorola. IEEE Softw. **14**(5), 75–81 (1997)
9. Yadav, H.B., Yadav, D.K.: A fuzzy logic based approach for phase-wise software defects prediction using software metrics. http://dx.doi.org/10.1016/j.infsof.2015.03.001
10. Promise Repository: http://promise.site.uottawa.ca/SERepository/datasets-page.html
11. Fenton, N.E., Neil, M., et al.: On the effectiveness of early life cycle defect prediction with Bayesian Nets. Empirical Softw. Eng. **13**(5), 499–37 (2008)
12. Vashisth, V., Lal, M., Sureshchander, G.S.: A framework for software defect prediction using neural networks

Heterogeneous Data Distortion for Privacy-Preserving SVM Classification

J. Hyma, P. Sanjay Varma, S. V. S. Nitish Kumar Gupta and R. Salini

Abstract The modern developments have made our lives' more digital. As they make our day-to-day activities simpler, they are in wide usage leading to accumulation of huge amounts of data. The analysis of such data will provide useful results for further advancements which can be obtained by the concept of Data Mining. These newly emerging digital exercises involve sharing of personal information to unknown sources, questioning the privacy of the individual. To avoid such situations, techniques of Privacy-Preserving Data Mining involving various data distortions are used. One such technique is Privacy-Preserving Data Classification, balancing both privacy and utility in classification aspects. In this work, the performance analysis of Classification using Support Vector Machine (SVM) on data sets that are heterogeneously distorted is evaluated.

1 Introduction

Fields like marketing, manufacturing, and research considers that analyzing people's data is very important. Day by day, bulk of data is being escalated with the use of modern gadgets. Analyzing these large data sets and obtaining patterns is an arduous task to accomplish. To achieve the computation at high speed, Data Mining [1] has always been the most efficient way. This process combines tools from statistics and artificial intelligence (such as neural networks and machine warehouses) that considers analysis of data from huge data sources.

J. Hyma (✉) · P. Sanjay Varma · S. V. S. Nitish Kumar Gupta · R. Salini
Department of CSE, GIT, GITAM University, Visakhapatnam, India
e-mail: jhyma.gitam@gmail.com

P. Sanjay Varma
e-mail: psanjayvarma5@gmail.com

S. V. S. Nitish Kumar Gupta
e-mail: nitish.somisetty@gmail.com

R. Salini
e-mail: rssalini7@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_49

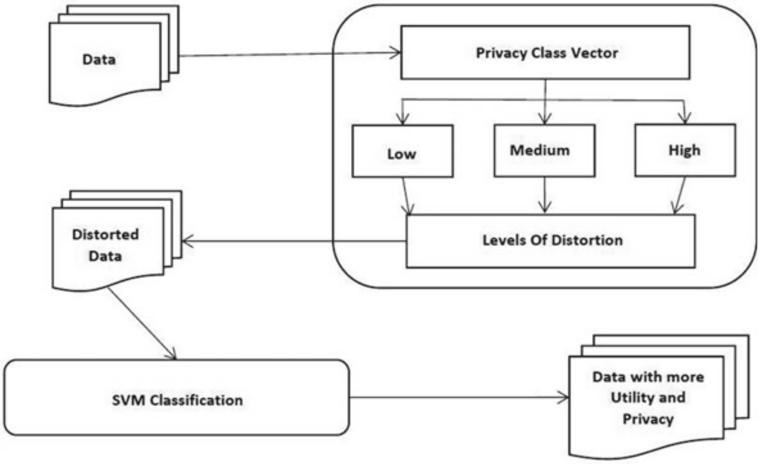


Fig. 1 Proposed model to privacy-preserving data mining

Data mining is useful for several purposes like predictions and recommendations. This process involves the usage of individual's personal information, i.e., one's private information is no more private. In countries like Europe under Directive 96/9/EC [2] of the European Parliament and of the Council of March 11, 1996, the usage of data without the permission of the data owner is a punishable offense.

Individuals might not be willing to provide/share their data due to privacy violations, which led to the concept of privacy preservation data mining. Privacy-Preserving Data Mining (PPDM) came into existence in order to overcome such violations, by the efficient utility of data along with privacy. As the main objective of PPDM is to build algorithms for transforming the original information in some way, so that the private data and private knowledge remain confidential even after the mining process. Various views like Correlation, Distortion, Classification, Normalization, and Clustering are proposed to achieve PPDM [3]. In the process of data mining acquiring the data involves three major role players, i.e., the Data sources, Data collectors, and Data Analysts [4].

Various views that ensure in preserving the privacy of the data owner are applied in the process of data mining. These views are implemented by the data collectors on the bulk data before it is fetched by the data analysts to maintain the owner's data privacy. The Data collectors choose an optimal method which ensures the maximum data utility by the data analysts and minimal privacy violation for the data owner [5]. Figure 1 gives an overview of the discussion.

In the paper, the required knowledge about Classification is provided in Sect. 2 and Privacy-Preserving Data Mining in Sect. 3, followed by Support Vector Machine (SVM) in Sect. 4 and Heterogeneous Data Distortion (HDD) in Sect. 5. In Sect. 6, the performance analysis of SVM on HDD is given and finally, conclusion of the works in Sect. 7.

2 Classification

Classification is a general process related to categorization. Classification involves analyzing the input data to develop an accurate description for each class using the features present in the data [6].

2.1 Need for Classification

Unlike the most data mining solutions, a classification usually comes with a degree of certainty. It might be the probability of the object belonging to the class or some other measure of how closely the object resembles other elements from that class.

Classification of future or unknown objects is a two-step process involving:

- Estimate accuracy of the model.
- If the accuracy is acceptable, use the model to classify data tuples whose class labels are not known [7].

2.2 Why Use Classification Techniques

The classification techniques benefit different industries through which they can easily identify the group and type to which the tuple belongs by classifying the data into classes [7]. Various applications of this include customer segmentation, business modeling, marketing, credit analysis, and biomedical and drug response modeling. Data mining is achieved by applying the classifications but the data owner is vulnerable to data privacy violation. Keeping this in mind, privacy-preserving data classification techniques are proposed.

3 Privacy-Preserving Data Classification

Generally, the data collector is the one who takes care of the privacy of the data owner by applying various privacy preservation techniques to the classification and provides the data to the data analyst. So, we need to use optimal privacy preservation techniques which exclude privacy violations. The main aim of privacy-preserving data classification [8] is to build classifiers that ensure data privacy of an individual. However, privacy and security restricts the free sharing of data which in turn make the data mining challenging. Privacy is very specific, i.e., the restriction imposed by the data user for their particular data [9].

The main privacy-preserving data classification techniques are classified into supervised and unsupervised learning techniques. Supervised learning technique has the knowledge of output and data is classified into a class or value unlike unsupervised learning technique losing an edge over supervised techniques. Support Vector Machine is one of the supervised classification technique is used because a probabilistic approach is used to classify the data and hence complex relations between data can be shown without the necessity of difficult transformation [10].

4 Support Vector Machine

Support Vector Machine simply SVM, is a supervised classification method generally used for regression analysis. SVM is a geometrical approach [11], as a result, the outcomes are more elaborated. It has a capability to produce outcomes with higher accuracy and good predictive analysis in comparison with higher classification algorithms [12]. So, it is better to accomplish purposes like facial expression recognition [13] and others as discussed in [14, 15].

4.1 Nonlinear Support Vector Machine

Nonlinear Support Vector Machine handles data sets that cannot be linearly classified like one-dimensional data [16]. Solving such kind of data sets involves methods such as Feature Mapping, Smallest Enclosing Hypersphere, Support Vector Regression, and Support Vector Ranking [17].

4.2 Linear Support Vector Machine

Linear Supporting Vector Machine simply LSVM, has the data sets which are linearly separable [12]. Here, data sets are considered as vectors and are plotted as points on the graph. Suppose, if there are n-data set elements x_1, x_2, x_3..., x_n, plotted on the graph as (x_i, y_i) where y_i is {1,−1}, i.e., $y_i = 1/-1$, which defines the class of x_i where each x_n is a p-dimensional vector. To classify each vector point, an ambient space is defined, called hyperplane [12, 18] which is an (p−1) dimensional plane. This hyperplane is considered as a set of points x which is represented in Eq. (1).

$$w.x - b = 0 \tag{1}$$

where w is a normal vector to the hyperplane and b is an offset. As the considered data is linearly separable, we select two parallel hyperplanes which is shown in Eqs. (2) and (3).

$$w.x - b = 1 \tag{2}$$

$$w.x - b = -1 \tag{3}$$

And satisfying the condition that any vector from the parallel hyperplanes should fall below or above the soft-margin and is represented in Eqs. (4) and (5) [14].

$$w.x - b \; >= 1 \tag{4}$$

$$w.x - b \; <= -1 \tag{5}$$

5 Heterogeneous Data Distortion

As mentioned before maintaining the data owner's privacy and using the data to the advantage of the data analyst should be the main objective of privacy-preserving data mining (PPDM). Privacy is a requirement which differs based on various factors. To achieve this PPDM, steps should be taken to ensure that there is a balance between the two key factors data utility and data privacy. One such factor can be a personal choice based on which, privacy of the data can be more even though it is not essential. In such criteria, the utility of the data decreases. This is the case in which the disclosure of data is done more by the data owner. In another case, the privacy of data can be decided solely by the data collector to increase the data utility thereby decreasing data privacy in some cases. So, there is a need to choose an alternative intermediate data distortion method from the previous work [5], i.e., Heterogeneous Data Distortion simply HDD [1].

In HDD, neither the privacy nor the utility is compromised. The data is distorted based on the chosen factors like [4] privacy choice vector, [4] domain privacy vector, and [4] correlation vector. Now, we are going to calculate the accuracy and error rate of Support Vector Machine technique on Heterogeneous Data Distortion considering the data sets from the previous work [19, 20].

6 Experimental Consideration

In this section, the result analysis performed on two different data sets is given in Tables 1 and 2. The Adult data set [19], Liver Data set [20] are rendered from the UCI Machine Repository.

Table 1 Comparison of accuracy, error rates of SVM technique

Data set	Applying SVM technique on	Accuracy	Error
Adult data set	Original data	85.21	14.78
	Perturbed data	80.58	19.41
	Original as Training set, Perturbed as Test Data	75.05	24.95
Liver data set	Original as Training set	93.89	6.1
	Perturbed as Training set	93.78	6.2
	Original as Test data, percentage split, cross-validation	65.35	34.64
	Perturbed as test data, percentage split, cross-validation	66.16	33.8

Table 2 Comparison of various parameters of data sets

Adult data set	**TP rate**	**FP rate**	**Precision**	**F-Measure**
Original data	0.852	0.332	0.845	0.845
Perturbed data	0.806	0.547	0.797	0.771
Original as Training set, Perturbed as Test Data	0.75	0.75	0.563	0.643
Liver data set	**TP rate**	**FP rate**	**Precision**	**F-Measure**
Original as Training set	0.988	0.016	0.988	0.988
Perturbed as Training set	0.988	0.012	0.988	0.988
Original as Test data, percentage split, cross-validation	0.654	0.494	0.656	0.655
Perturbed as test data, percentage split, cross-validation	0.653	0.482	0.654	0.655

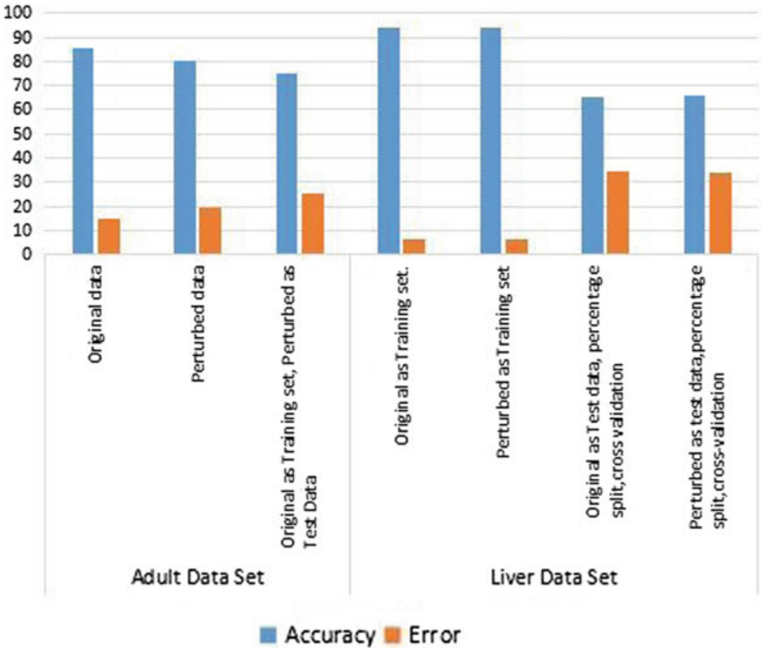


Fig. 2 Comparison plot of accuracy and error rates

The distorted data accuracy along with the original data set is measured. Using WEKA tool [21], the accuracy results with the SVM Classification technique are obtained. The results in Table 1 which are graphically plotted in Figs. 2, 3, and 4 show that the accuracy obtained by the distortion method is almost similar to the model developed using original data set.

7 Conclusion

The given work proposed a performance analysis of SVM classification technique on Heterogeneously Distorted Data. The heterogeneously distorted data is obtained by applying the HDD technique proposed in [4]. We discussed about the need for classification, uses and various types of classification techniques, privacy preservation data

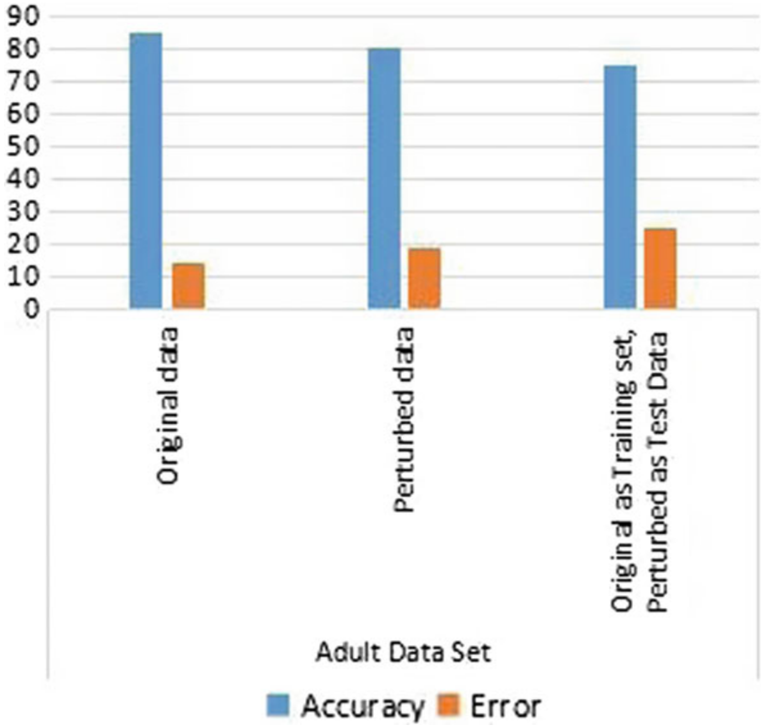


Fig. 3 Comparison plot of adult data set

classification. To hold the privacy and utilize the data of the user, we apply SVM on the heterogeneously distorted data. The results are deduced by comparing the accuracy rates, error rates of the original data and HDD data by applying SVM. From the results obtained using SVM, it proves to be a satisfying classification method which can be used in Privacy-preserving data mining to achieve good results while providing a proper balance in both privacy and utility.

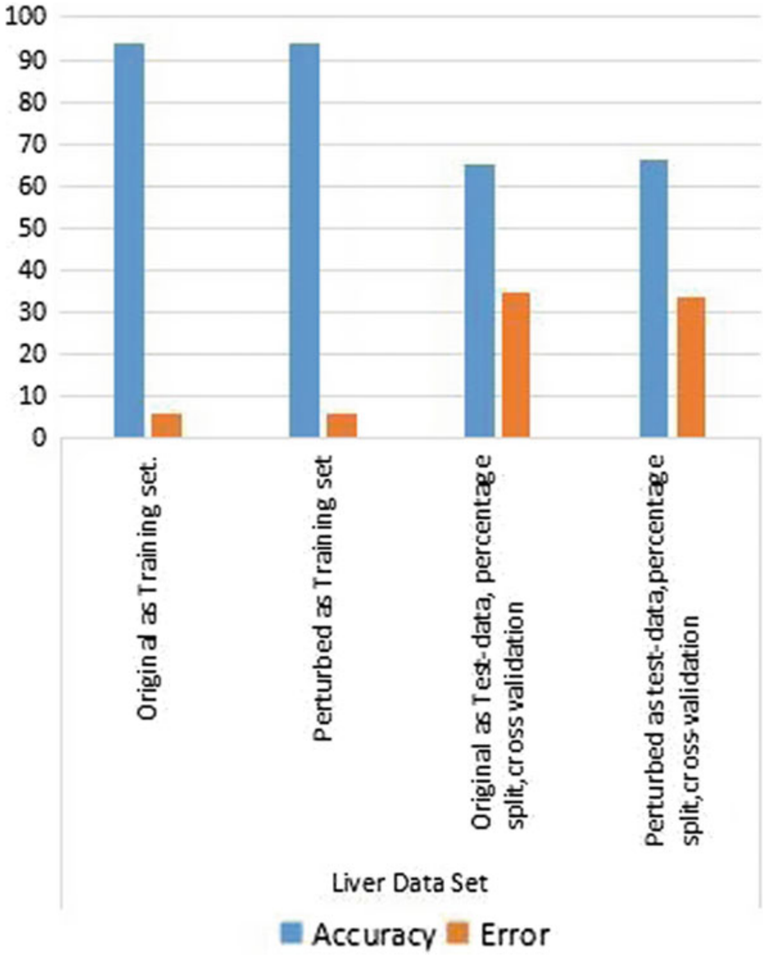


Fig. 4 Comparison plot of liver data set

References

1. Data Mining Curriculum: ACM SIGKDD. 2006-04-30. Retrieved 27 Jan 2014
2. Council Decision of 16 March 2000 on the approval, on behalf of the European Community, of the WIPO Copyright Treaty and the WIPO Performances and Phonograms Treaty (2000/278/EC), OJ no. L089 of 2000-04-11, pp. 6–7
3. Aldeen, Y.A.A.S., Salleh, M., Razzaque, M.A.: A comprehensive review on privacy preserving data mining. SpringerPlus **4**, 694 (2015)
4. Hyma, J., Prasad Reddy, P.V.G.D., Damodaram, A.: Performance analysis of heterogeneous data normalization with a new privacy metric. Int. J. Comput. Sci. Inf. Secur. (IJCSIS) **14**(6) (2016)
5. Hyma, J., Prasad Reddy, P.V.G.D., Damodaram, A.: A study of correlation impact on privacy preserving data mining. Int. J. Comput. Appl. (0975–8887) **129**(15) (2015)

6. Zhang, G.: Neural networks for classification: a survey. IEEE Trans. Syst. Man Cybern. Part C **30**(4), 451–462 (2000)
7. Gupta, M., Aggarwal, N.: Classification techniques analysis. In: NCCI 2010 -National Conference on Computational Instrumentation CSIO Chandigarh, India, 19–20 Mar 2010
8. Zhang, N., Wang, S., Zhao, W.: A new scheme on privacy-preserving data classification. In: KDD‘05 Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, ACM New York, NY, USA ©2005 table of contents. Aug 2005. ISBN:1-59593-135-X. https://doi.org/10.1145/1081870.1081913
9. Shah, A., Gulati, R.: Privacy preserving data mining: techniques, classification and implications—a survey. Int. J. Comput. Appl. (0975–8887) **137**(12) (2016)
10. Lee, Y.: Support vector machines for classification: a statistical portrait. In: Bang, H., Zhou, X., van Epps, H., Mazumdar, M. (eds.) Statistical Methods in Molecular Biology. Methods in Molecular Biology (Methods and Protocols), vol. 620. Humana Press, Totowa, NJ (2010)
11. Vaidya, J., Yu, H., Jiang, X.: Privacy-preserving SVM classification. Published online: 24 Mar 2007, Springer-Verlag London Limited (2007)
12. Srivastava, D.K., Bhambhu, L.: Data Classification Using Support Vector Machine. 2005–2009 JATIT. All rights reserved
13. Michel, P., Kaliouby, R.E.: Real time facial expression recognition in video using Support Vector machines. In: Proceedings of ICMI’03, pp. 258–264 (2003)
14. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene Selection for cancer classification using support vector machines. Mach. Learn. **46**(1–3), 389–422 (1995)
15. Fradkin, D., Muchnik, I.: Support vector machines for classification. In: DIMACS Series in Discrete Mathematical and Theoretical Computer Science, Mathematical Subject classification, 62H30 (2000)
16. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
17. Aizerman, M.A., Braverman, E.M., Rozonoer, L.I.: Theoretical foundations of the potential function method in pattern recognition learning. Autom. Remote Control **25**, 821–837 (1964)
18. Berwick, R.: An Idiot’s guide to Support vector machines (SVMs). Retrieved on Oct 2003. mit.edu
19. https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data
20. https://github.com/ThachNgocTran/PredictLiverPatientWithRandomForestAndLogisticRegression/blob/master/Indian%20Liver%20Patient%20Dataset%20(ILPD).csv
21. Frank, E., Hall, M.A., Witten, I.H.: The WEKA Workbench. Online Appendix for “Data Mining: Practical Machine Learning Tools and Techniques”, 4th edn. Morgan Kaufmann (2016)

Microarray Analysis Using Multiple Feature Data Clustering Algorithms

B. SivaLakshmi and N. Nagamalleswara Rao

Abstract By using Microarray Technology, in a single experiment, one can study the function of thousands of genes in parallel. Microarrays are used in various applications like disease diagnosis, drug discovery, and biomedical research. A Microarray image contains thousands of spots and each of the spot contains multiple copies of single DNA sequence. The analysis of microarray image is done in three stages: gridding, segmentation, and information extraction. The microarray image analysis takes the spot intensity data as input and produces the spot metrics as output which are used in classification and identification of differently expressed genes. The intensity of each spot indicates the expression level of the particular gene. This paper presents multiple feature clustering algorithms which extend the single feature (pixel intensity) clustering algorithms for segmentation of microarray image. The qualitative and quantitative results show that multiple feature clustering algorithms are more efficient than single feature clustering algorithms in segmenting the spot area, thus producing more accurate expression ratio.

1 Introduction

The workflow of microarray image analysis was separated into four stages [1].

I. *Image merging*, is the construction of the combined eight-bit image from intensity measurements of both red (Cy5) and green (Cy3) fluorescent dye, that is computationally efficient in doing subsequent gridding and segmentation steps. The combined image I is obtained by using some arbitrary function f, i.e., $I(i, j) = f(R(i, j), G(i, j)$ where R is an image corresponding to red channel and G is an image corresponding to green channel.

B. SivaLakshmi (✉)
Acharya Nagarjuna University, Vijayawada, AP, India
e-mail: bslakshmi85@gmail.com

N. Nagamalleswara Rao
Department of IT, RVR and JC College of Engineering, Guntur, AP, India

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_50

II. *Gridding* [2], is the mechanism of identification of location of the gene spots in the image without any overlapping. *Sub-gridding* refers to finding the block index corresponding to a spot on the microarray image, while *spot-detection*, is finding the location *(i, j)* of a specified spot in that indexed block.

III. *Segmentation* [3], is the problem of classifying the pixels of image into a set of nonoverlapping regions based on specific criteria. In microarray image, the pixels can be classified into spot, background or noise.

IV. *Information Extraction* [4], includes the calculation of metrics such as Means and Medians, Standard deviation, Diameter, Expression Ratio, etc., in the region of every gene spot on the microarray image. The expression ratio measures the transcription abundance between the two sample gene sets. The positive or negative expression ratio indicates the overexpression or underexpression between the control and treatment genes.

2 Enhancement and Gridding

If the contrast of the microarray image is low, the quality of the edges extracted from the image will be poor. This edge information is the primary source for automatic gridding of microarray image [5]. The quality of the spot edges can be improved by applying GA based contrast enhancement algorithm [6] to the original image prior to the computation of Gridding. Gridding is the process of dividing the microarray image into blocks (sub-gridding) and each block again divided into subblock (spot-detection). The final sub-block contains a single spot and having only two regions spot and background. Existing algorithms for gridding are semi-automatic in nature requiring several parameters such as size of spot, number of rows of spots, number of columns of spot, etc. In this paper, a fully automatic gridding algorithm designed in [7, 8] is used for sub-gridding and spot-detection.

3 Segmentation

Many microarray image segmentation approaches have been proposed in literature. Fixed circle segmentation [9], Adaptive circle Segmentation Technique [10], Seeded region growing methods [11] and clustering algorithms [12] are the methods that deal with microarray image segmentation problem. This paper mainly focuses on clustering algorithms. These algorithms have the advantages that they are not restricted to a particular spot size and shape, does not require an initial state of pixels and no need of post-processing. These algorithms have been developed based on the information about the intensities of the pixels only (one feature). But in the microarray image segmentation problem, not only the pixel intensity but also the distance of pixel from the center of the spot and median of intensity of a certain number of surrounding pixels influence the result of clustering. In this paper, multiple feature clustering algorithms are proposed, which utilizes more than one feature. The qualitative and quantitative results show that multiple feature clustering algorithms have segmented the image better than single feature clustering algorithms.

4 Single Feature Clustering Algorithms

***K*-means Algorithm**: The k-means clustering algorithm [13] for segmentation of microarray image is described as follows:

1. Randomly consider K initial clusters $\{C_1, C_2,\ldots\ldots,C_k\}$ from the $m*n$ image pixels $\{I_1, I_2, I_3,\ldots\ldots,I_{m*n}\}$.
2. Assign each pixel to the cluster C_j $\{j = 1,2,\ldots K\}$ if it satisfies the following condition:

$$D(I_i, C_j) < D(I_i, C_q),\ q = 1, 2, \ldots, K \\ j \neq q \tag{1}$$

   where $D(.\,,.)$ is the Euclidean distance measure between two values.
3. Find new cluster centroid as follows:

$$C_i^{\wedge} = \frac{1}{n_i} \sum_{I_j \in C_i} I_j,\ i = 1, 2, \ldots K \tag{2}$$

   where n_i is the number of pixels belonging to cluster C_i.
4. If

$$C_i^{\wedge} = C_i,\ i = 1, 2, \ldots K \tag{3}$$

   Then stop.
   Else continue from step 2.

***K*-medoids Algorithm**: The *k*-medoids clustering algorithm [14] for segmentation of microarray image is described as follows:

1. Randomly consider K initial medoids $\{M_1, M_2,\ldots\ldots, M_k\}$ for the clusters $\{C_1, C_2,\ldots\ldots,C_k\}$ from the $m*n$ image pixels $\{I_1, I_2, I_3,\ldots\ldots, I_{m*n}\}$.
   A cluster medoid is a point that is located centrally within the cluster. It is the point that has minimum sum of distances to other points of the cluster.
2. Assign each pixel to the cluster C_j $\{j = 1,2,\ldots.K\}$ if it satisfies the following condition:

$$D(I_i, M_j) < D(I_i, M_q),\ q = 1, 2, \ldots, K \\ j \neq q \tag{4}$$

where $D(.,.)$ denotes the dissimilarity measure.

3. Find new medoids $M_i^\wedge$ belonging to clusters C_i, $i = 1, 2, \ldots K$. It is the pixel value with minimum total dissimilarity to all other points.
4. If

$$M_i^\wedge = M_i,\ i = 1, 2, \ldots K \tag{5}$$

Then stop.
Else continue from step 2.

***K*-modes Algorithm**: The *k*-modes clustering algorithm [15] is similar to *k*-medoids where modes are used instead of medoids. The *K*-modes algorithm for segmentation of microarray image is described as follows:

1. Randomly consider K initial modes {MO_1, MO_2,……,MO_k} for the clusters {C_1, C_2,……,C_k} from the m*n image pixels {I_1, I_2, I_3,……, I_{m*n}}.
   A mode is a pixel value within a cluster that is repeated more often than any other value.
2. Assign each pixel to the cluster C_j {j = 1, 2,…..K}, if it satisfies the following condition:

$$\begin{aligned}&D(I_i, MO_j) < D(I_i, MO_q),\ q = 1, 2, \ldots, K\\ &j \neq q\end{aligned} \tag{6}$$

where $D(.,.)$ denotes the dissimilarity measure.

3. Find new medoids $MO_i^\wedge$ belonging to clusters C_i, $i = 1, 2, \ldots K$.
4. If

$$MO_i^\wedge = MO_i,\ i = 1, 2, \ldots K \tag{7}$$

Then stop.
Else continue from step 2.

Fuzzy *c*-means Algorithm: The Fuzzy *c*-means clustering algorithm [16] for segmentation of microarray image is described as follows:

1. Randomly consider K initial clusters {C_1, C_2,……,C_k} from the $m*n$ image pixels {I_1, I_2, I_3,……,I_{m*n}}.
2. The membership matrix u_{ij} is initialized with value from 0 to 1 and value of m = 2. The summation of pixel memberships representing particular cluster should be equal to 1.
3. Assign each pixel to the cluster C_j {j = 1, 2,…K}, if it satisfies the following condition:

$$\begin{aligned}&u_{ij}^m D(I_i, C_j) < u_{iq}^m D(I_i, C_q),\ q = 1, 2, \ldots, K\\ &j \neq q\end{aligned} \tag{8}$$

where $D(.\,,.)$ is the Euclidean distance measure between two values.

4. Find new membership and cluster centroid values as follows: $u_{ik} = \frac{1}{\sum_{j=1}^{K}\left(\frac{D(C_i, I_k)}{D\left(C_j, I_k\right)}\right)^{\frac{1}{m-1}}}$, for $1 \le i \le K$ u_{ik} denotes the *k*th object in the *i*th cluster.

$$C_i^{\wedge} = \frac{\sum_{j=1}^{n} u_{ij}^{m} I_j}{\sum_{j=1}^{n} u_{ij}^{m}} \tag{9}$$

where n is the number of pixels belonging to cluster C_i.

5. Continue 2–3 until each object is assigned to the cluster that has maximum membership [17].

Fuzzy *K*-medoids Algorithm Segmentation: The Fuzzy *k*-medoids [18] clustering algorithm is the extension of fuzzy K-means algorithm replacing means by medoids.

5 Multiple Feature Clustering Algorithm

The clustering algorithms used for microarray image segmentation are based on the information about the intensities of the pixels only. But in microarray image segmentation, the position of the pixel and median value of surrounding pixels also influences the result of clustering and subsequently that leads to segmentation. Based on this observation, multiple feature clustering algorithms is developed for segmentation of microarray images. To apply clustering algorithms on a single spot, we take all the pixels that are contained in the spot are, which is obtained after gridding process, and create a dataset $D = \{x_1, x_2, x_3, x_4, x_5, \ldots\ldots, x_n\}$, where $x_i = [x_i^{(1)}, x_i^{(2)}, x_i^{(3)}]$ is a three-dimensional vector that represents the *i*th pixel in the spot region. We use three features, which are defined as follows:

$x_i^{(1)}$: Represents the pixel intensity value.

$x_i^{(2)}$: Represents the distance from pixel to the weighted center of the spot region. The spot center is calculated as follows:

1. Apply edge detection to the spot region image using canny method.
2. Perform flood-fill operation on the edge image using imfill method.
3. Obtain label matrix that contain labels for the 8-connected objects using bwlabel function.
4. Calculate the centroid of each labeled region (connected component) using region props method.

The coordinates of the *i*th pixel are represented by two-dimensional vector $p_i = [p_x, p_y]^{t}$, then

$x_i^{(2)} = ||p_i\text{-}c||$, where c represents the weighted center of the spot region obtained as

$$c = \frac{1}{\sum_{i=1}^{n} x_i^{(1)}} \sum_{i=1}^{n} x_i^{(1)} p_i. \tag{10}$$

$x_i^{(3)}$: Represents the mean, median, or variance of the intensity of surrounding pixels. Here we have used variance of surrounding pixels.

For each pixel in the spot region, once the features are obtained forming the dataset D, then the clustering algorithms is applied.

6 Experimental Results

Qualitative Analysis: The proposed clustering algorithm is performed on two microarray images drawn from the standard microarray database corresponds to breast category a CGH tumor tissue. Image 1 consists of a total of 38,808 pixels and Image 2 consists of 64,880 pixels. Gridding is performed on the input images by the method proposed in [13], to segment the image into compartments, where each compartment is having only one spot region and background. The gridding output is shown in Fig. 1. After gridding the image into compartments, such that each compartment is having single spot and background, compartment No. 1 from image 1 and compartment No. 8 from image 2 are extracted. The image compartments are segmented using multiple feature clustering algorithms. The segmentation result of multiple feature fuzzy c-means is shown in Table 1. Table 1 shows the quantitative evaluations of clustering algorithms using MSE. The results confirm that multiple feature fuzzy *c*-means algorithm produces the lowest MSE value for segmenting the microarray image.

Table 1 MSE values

Method	Single feature clustering		Multiple feature clustering	
	Compartment No. 1	Compartment No. 8	Compartment No. 1	Compartment No. 8
K-means	96.2	95.8	95.4	93.8
K-medians	95.1	94.2	94.2	92.2
K-mode	95.2	94.3	94.4	92.6
Fuzzy *c*-means	90.7	90.2	88.7	88.5
Fuzzy *K*-medoids	91.8	91.4	89.2	89.9

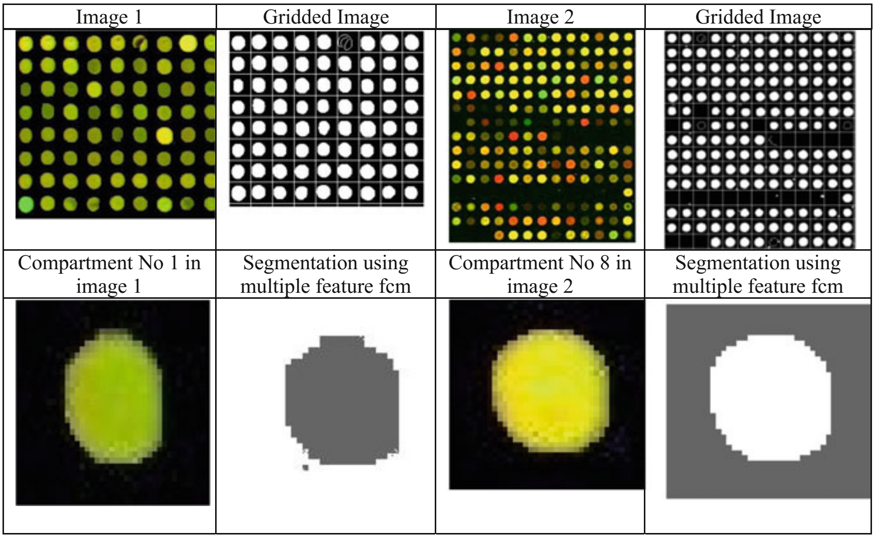


Image 1
Gridded Image
Image 2
Gridded Image
Compartment No 1 in image 1
Segmentation using multiple feature fcm
Compartment No 8 in image 2
Segmentation using multiple feature fcm

Fig. 1 Gridding and segmentation results

7 Conclusions

Microarray technology provides simultaneous monitoring of thousands of gene expression levels. The main steps in microarray image analysis are gridding, segmentation and information extraction. Multiple feature clustering algorithms are used for segmentation of microarray image. These algorithms are extensions to the *k*-means clustering algorithm. These algorithms have been developed based on the information about the intensities of the pixels only (one feature). But in the microarray image segmentation problem, not only the pixel intensity, but also the distance of pixel from the center of the spot and median of intensity of a certain number of surrounding pixels influence the result of clustering. Based on these observations, multiple feature clustering algorithms are presented in this paper. Out of these algorithms, multiple feature fuzzy *c*-means clustering algorithm produces better segmentation result. Log ratio of R/G gives the abundance of expression level of the corresponding gene.

References

1. Schena, M., Shalon, D., Davis, R.W., Brown, P.O.: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science **270**, 467–470 (1995)
2. Chen, W.B., Zhang, C., Liu, W.L.: An automated gridding and segmentation method for cDNA microarray image analysis. In: 19th IEEE Symposium on Computer-Based Medical Systems
3. Tsai, T.H., Yang, C.P., Tsai, W.C., Chen, P.H.: Error Reduction on Automatic Segmentation in Microarray Image. IEEE (2007)

4. Harikiran, J., et al.: Vector filtering techniques for impulse noise reduction with application to microarray images. Int. J. Appl. Eng. Res. **10**(3), 7181–7193 (2015)
5. Harikiran, J., Raghu, A., Lakshmi, P.V., Kiran Kumar, R.: Edge detection using mathematical morphology for gridding of microarray image. Int. J. Adv. Res. Comput. Sci. **3**(2), 172–176 (2012)
6. Sivalakshmi, B., Nagamalleswara rao, N.: Microarray image analysis using genetic algorithm. IAES Indonesian J. Electr. Eng. Comput. Sci. **4**(3), 561–567 (2016)
7. Harikiran, J., Lakshmi, P.V., Kirankumar, R.: Automatic gridding method for microarray images. J. Theor. Appl. Inf. Technol. **65**(1), 235–241 (2014)
8. Harikiran, J., Ramakrishna, D., Avinash, B., Lakshmi, P.V., Kiran Kumar, R.: A new method of gridding for spot detection in microarray images. Comput. Eng. Intell. Syst. **5**(3), 25–33 (2014)
9. Eisen, M.: ScanAlyze User's manual (1999)
10. Buhler, J., Ideker, T., Haynor, D.: Dapple: improved techniques for finding spots on DMA microarray images. Tech. Rep. UWTR 2000-08-05, University of Washington (2000)
11. Adams, R., Bischof, L.: Seeded region growing. IEEE Trans. Pattern Anal. Mach. Intell. **16**(6), 641–647 (1994)
12. RamaKrishna, D., Harikiran, J., et al.: Various versions of K-means clustering algorithm for segmentation of microarray image. Int. J. Electron. Commun. Comput. Eng. **4**(1), 1554–1558 (2012)
13. Harikiran, J., Lakshmi, P.V., Kiran Kumar, R.: Fast clustering algorithms for segmentation of microarray images. Int. J. Sci. Eng. Res. **5**(10), 569–574 (2014)
14. Anirban, M., et al.: Multiobjective genetic algorithm based fuzzy clustering of categorical attributes. IEEE Trans. Evol. Comput. **13**(5), 991–1005 (2009)
15. Kiran Kumar, R., Saichandana B., et.al.: Dimensionality reduction and classification of hyperspectral images using genetic algorithm. IAES Indonesian J. Electr. Eng. Comput. Sci. **3**(3), 503–511 (2016)
16. Harikiran J., et.al.: Fuzzy C-means with bi-dimensional empirical mode decomposition for segmentation of microarray image. Int. J. Comput. Sci. Issues **9**(5), Number 3, 273–279 (2012)
17. Saichandana, B., et al.: Image fusion in hyperspectral image classification using genetic algorithm. IAES Indonesian J. Electr. Eng. Comput. Sci. **2**(3), 703–711 (2016)
18. Saichandana, B., et al.: Clustering algorithm combined with hill climbing for classification of remote sensing image. IAES Int. J. Electr. Comput. Eng. **4**(6), 923–930 (2014)

Privacy Preserving Data Clustering Using a Heterogeneous Data Distortion

Padala Preethi, Kintali Pavan Kumar, Mohammed Reezwan Ullhaq, Anantapalli Naveen and Hyma Janapana

Abstract Modern age computation leads to huge amount of data. The whole data is analysed using data mining. In return, it made its path to disruption of the privacy of data owners. In order to achieve privacy on data we use Privacy Preserving Data Mining (PPDM). But when the privacy is maintained the data utility is decreased and vice versa. So, in order to maintain a balance in both privacy and data utility, Privacy Preserving Data Clustering (PPDC) using a Heterogeneous data distortion is introduced. In this article both original and perturbed data are analysed using K-means and density based clustering techniques and the results are compared to show the balance between privacy and utility of the data.

1 Introduction

There is an enormous amount of data present in the Information Industry. Until the data is not transformed into useful information it cannot be used properly. It is important to analyse this data and obtain appropriate information. In order to analyse the huge data in an effective and efficient way, Data Mining [1, 2] has been the best possible technique. This process includes inspecting, cleansing, transforming and modelling data to achieve useful information.

P. Preethi
NIT, Surathkal, India

K. P. Kumar · M. R. Ullhaq · A. Naveen (✉) · H. Janapana
Department of Computer Science, Gandhi Institute of Technology, Visakhapatnam, India
e-mail: anaveen2000@gmail.com

K. P. Kumar
e-mail: k.pavankumar1061@gmail.com

M. R. Ullhaq
e-mail: mdreezwan@gmail.com

H. Janapana
e-mail: jhyma.gitam@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_51

Data mining is used for many purposes. For example, it is used for developing smarter marketing campaigns and also for the accurate prediction of customer loyalty, in this process data mining techniques extract individual's personal information, i.e., individual's private data is being exposed. In some countries, this act is a punishable offence [3, 4].

Some individual might not be willing to share their sensitive information due to some privacy concerns. The problems regarding privacy in data mining are solved using a new technique called as privacy preserving data mining (PPDM) [5, 6].

PPDM research usually follows hiding of the sensitive data like id, name, age, etc. These sensitive datasets are changed, blocked, or cut down from the database, in order to protect the user's data by denying the access to the original data to other parties. Data is also encrypted before it is released or shared for computations; so that no party knows anything about the sensitive data excluding the results given for their own inputs.

The final objective of PPDM is to allow a person to extract appropriate information from an enormous amount of data by introducing powerful algorithms and at the same time avoiding the disclosure of sensitive data and information. A lot of analysis has been done in privacy preserving data mining [5] based on randomization, perturbation and Anonymity including l-diversity and K-anonymity.

The existing PPDM techniques [7, 8] are applying a universal distortion method to protect individual's privacy by ignoring the mandatory privacy parameters. However, this distortion is not sufficient for data privacy. So, there is an immense need of modifying the data heterogeneously and thus it better preserves the data utility and also owner's privacy. Clustering [9] is a separation of data into groups of similar objects. Each group is called a cluster. And it consists of objects that are similar to each other, objects of one group are different from the objects of other groups.

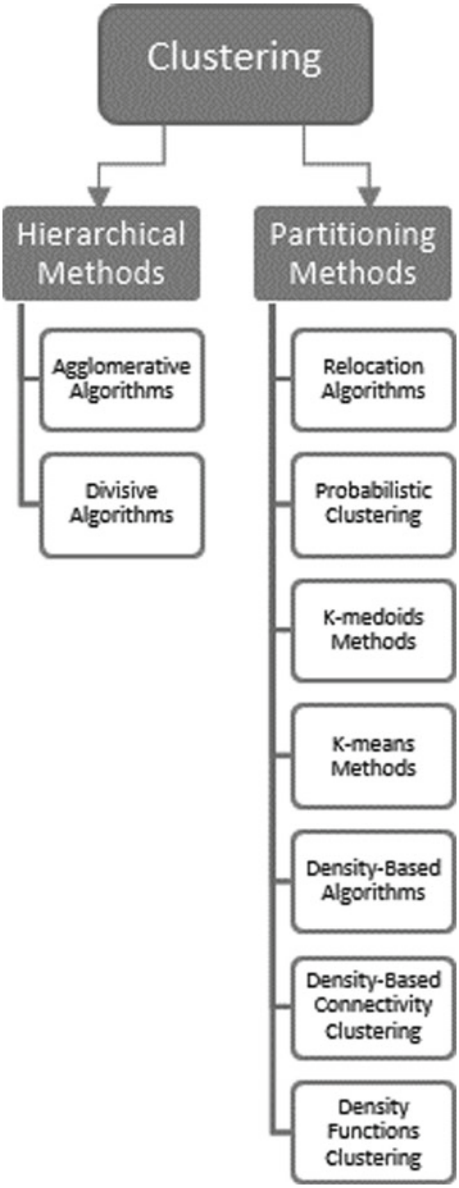
2 Clustering

Clustering is the process of dividing the enormous amount of data into groups called clusters which contains the related objects of data. In these clusters it only has the objects which are related to each other. And objects contained in one group are dissimilar from the others (Fig. 1).

2.1 Need for Clustering

There is a lot of raw information in the market which has to be used for several purposes. Now to make this raw information useful we use data mining. Data mining transforms the information which is derived from data sets into a simple understandable format.

Fig. 1 Classification of clustering



In this extracted information, we can make more useful datasets such as all the similar objects are put into one data set. This can be done using clustering. Using clustering we filter the information by putting each of them into a dataset which has similar information. These datasets can be used for Exploratory data mining can be

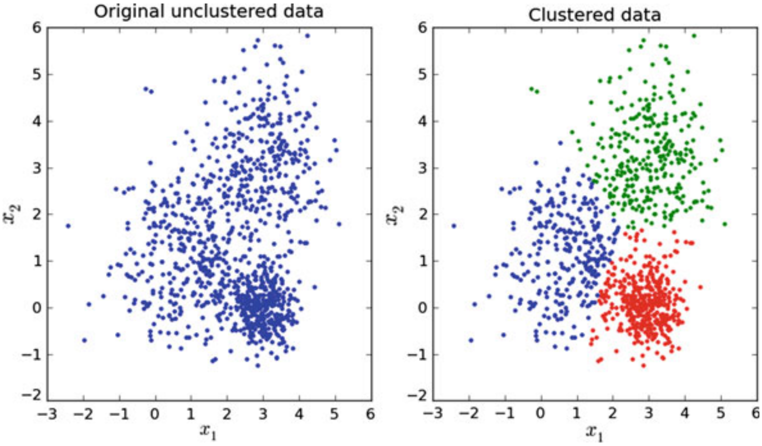


Fig. 2 Visual example of clustering

done using these data sets and also it is used as a common technique for statistical data analysis.

Clustering recognizes the groups of related datasets. And to explore other relationships these datasets are used.

Now let us take an example of an online shopping company which sells a wide range of products. If we need to discontinue some of the products which are low on sales we need to know about the sales of all the products. This can be done using data mining but is a huge process. If we use clustering in this case it simply clusters the products on the basis of their sales and separates. But here the individual data is vulnerable to data privacy violation. Keeping this in mind privacy preserving data clustering techniques are proposed (Fig. 2).

3 Privacy Preserving Data Clustering

Data mining is made easier by applying appropriate clustering techniques [10, 11]. As, on one end individual information is very crucial for business organizations to improve their business and on the other end, privacy of the individual is overstepped.

To prevent this from happening we must introduce the concepts of privacy preserving on clustering. There are numerous methods [12] used for preserving the privacy of the data while clustering. Some of these methods are vector quantization, fuzzy, Density based approach, etc. These methods increase the processing time and also decreases the data utility.

The main moto is to provide privacy along with good utility rate. For this purpose, the concept of K-means and Density based algorithm's is being used which is discussed in the next session.

3.1 K-Means Clustering Algorithm

K-means clustering [13, 14] is a process of vector quantization. The given data is formed into clusters by dividing the n observations into k clusters where each observation belongs to the cluster containing the nearest mean. Here the quality of each cluster can be measured using the objective function which is the sum of the squares of the distances from each point to the centroid of the cluster to which it is assigned.

Using the above K-means algorithm we performed clustering on adult and income data sets and the Utility percentages are calculated for original and perturbed data sets. The results are mentioned in the below sections.

3.2 Density-Based Clustering

Density-based clustering algorithm [14, 15] is used when we need to deal with large data sets associated with noise. Based on the density of points clusters are identified. Regions with high density depicts the presence of clusters whereas regions with low density indicates presence of noise and outliers. Concepts of density reach ability and density connectivity are implemented here.

In density-based clustering we have two types of points they are core points and non-core points a point is said to be a core point if it forms a cluster with points that are reachable from it reachable. Whereas non-core point cannot be reached from a point All points in a cluster are mutually connected. If a point is density reachable then it is the part of cluster as well.

Using the above k-means algorithm we performed clustering on adult and income data sets and the Utility percentages are calculated for original and perturbed data sets. The results are mentioned in the below sections.

Even though we used perturbed data there is a negligible change in utility but using density based algorithm gives rise to more utility when compared to k-means algorithm which is shown in the below result section.

4 Heterogeneous Data Distortion

The main objective of the Privacy Preserving Data Mining (PPDM) is to maintain the owner's privacy and also the data analyst must be benefited by using the data provided.

So, to achieve this we make sure that there is a balance between both the key factors, that is data utility and data privacy. There are two cases, in which one of them is that the privacy of data is decided by the data owner. So, the data can have more privacy even though it is not useful. In this situation, the data utility decreases.

Another case is the privacy of data being decided by the data collector. Here the data collector tries to decrease the privacy in order to increase the utility of data. So, to avoid this issue we need to choose an alternative data distortion method from previous work, i.e., Heterogeneous Data Distortion (HDD) [16].

In HDD, the distortion is done in such a way that each individual data contains various levels of sensitivity [16]. The levels of sensitivity are chosen according to the owner's choice and also with respect to the value present in the data source and various other factors.

4.1 Heterogeneous Privacy Preserving Data Clustering

Figure 3 depicts the flow of data in a heterogeneous privacy preserving data clustering model. Initially data along with heterogeneous constrains is retrieved from user and then data is mapped to its corresponding privacy class. Every privacy class has its own extent of distortion level, the data is distorted accordingly. Then that data is clustered by Density based approach and also using K-means approach. Based on the values obtained performance analysis is done.

5 Experimental Results

In this section, the analysis is performed on two different datasets, i.e., Adult dataset [17] and Income dataset [18] (dataset details are given in Table 1) using two different clustering techniques, i.e., simple k-means and density-based clustering. The result analysis is given in Tables 2 and 3.

Table 1 Datasets details

Data set	Attributes	Instances	Classes
Adult data set	6	29,262	2
Income data set	10	100	2

Table 2 Comparison of incorrectly clustered instances and utility percentage in K-means technique

Data sets	Applying k-means technique on	Incorrectly clustered instances (x) (%)	Utility percentage ($100 - x$) (%)
Adult data set	Original data	46.39	53.61
	Perturbed data	47.83	52.17
Income data set	Original data	43.43	56.57
	Perturbed data	45.47	54.53

Fig. 3 Data flow model of heterogeneous privacy preserving data clustering

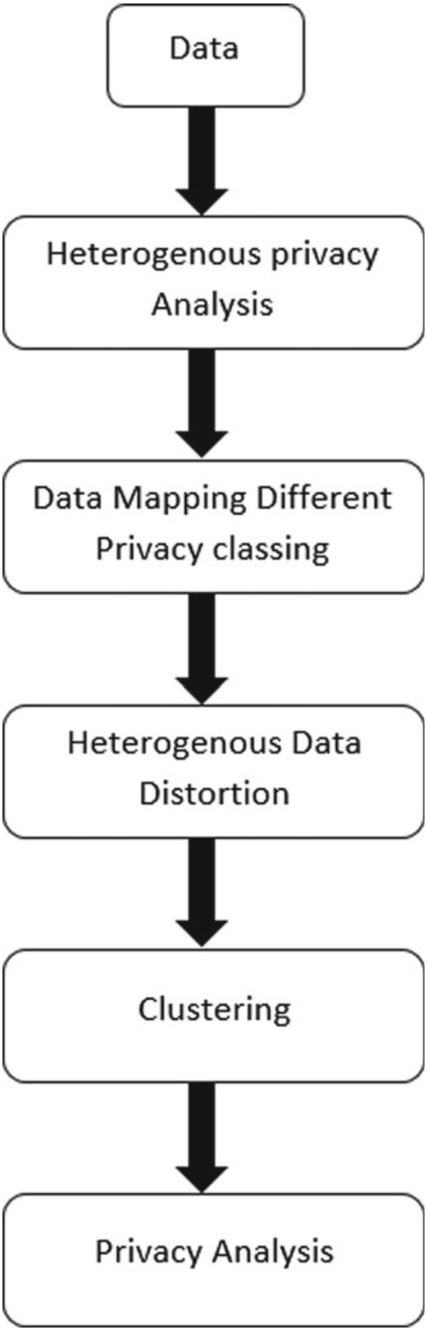

Data
Heterogenous privacy Analysis
Data Mapping Different Privacy classing
Heterogenous Data Distortion
Clustering
Privacy Analysis

Table 3 Comparison of incorrectly clustered instances and utility percentage in density based technique

Data sets	Applying density based technique on	Incorrectly clustered instances (x) (%)	Utility percentage ($100 - x$) (%)
Adult data set	Original data	44.47	55.53
	Perturbed data	45.16	54.84
Income data set	Original data	42.54	57.46
	Perturbed data	44.37	55.63

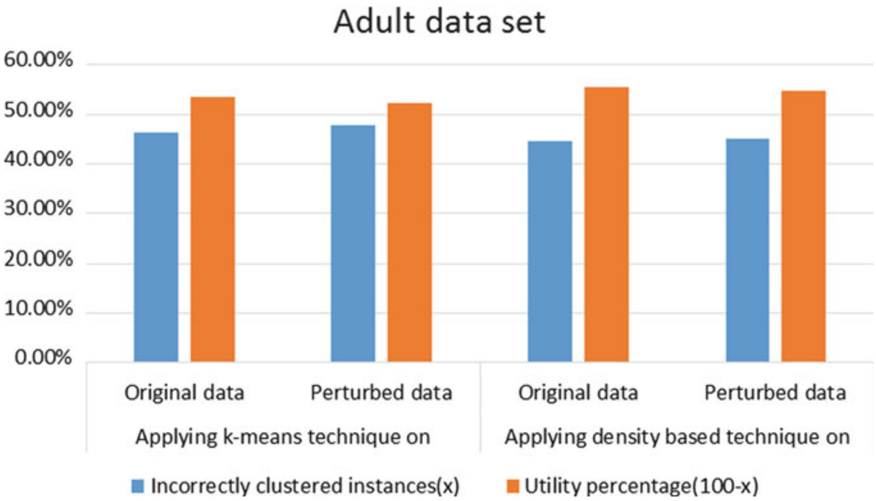


Fig. 4 Graphical representation of adult data sets

Tables 1 and 2 shows the comparison original data along with perturbated data set which is measured using weka tool [19], the results with k-means and density based are obtained. The results obtained shows the incorrectly clustered instances and Utility percentage.

Figure 4 illustrates the incorrectly clustered instances (x) and the utility percentage ($100 - x$) of the adult data set. It shows the values of original data and perturbed data by applying k-means technique and density based technique.

Figure 5 illustrates the incorrectly clustered instances (x) and the utility percentage ($100 - x$) of the income data set. It shows the values of original data and perturbed data by applying k-means technique and density based technique.

Figure 6 illustrates the incorrectly clustered instances (x) and the utility percentage ($100 - x$) of both adult data set and income data set. It shows the values of original data and perturbed data by applying k-means technique and density based technique.

K-means being the basic clustering algorithm we obtain low utility rates, to improve the utility rates a better clustering technique should be applied so in order

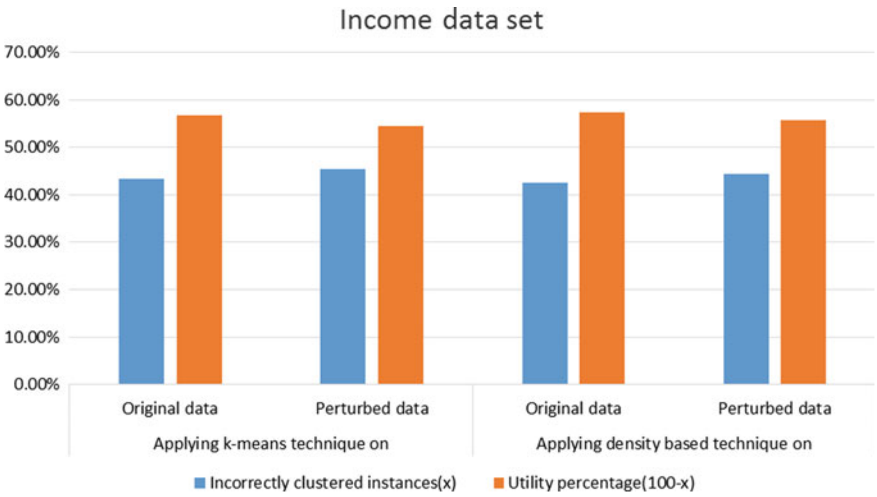


Fig. 5 Graphical representation of income data sets

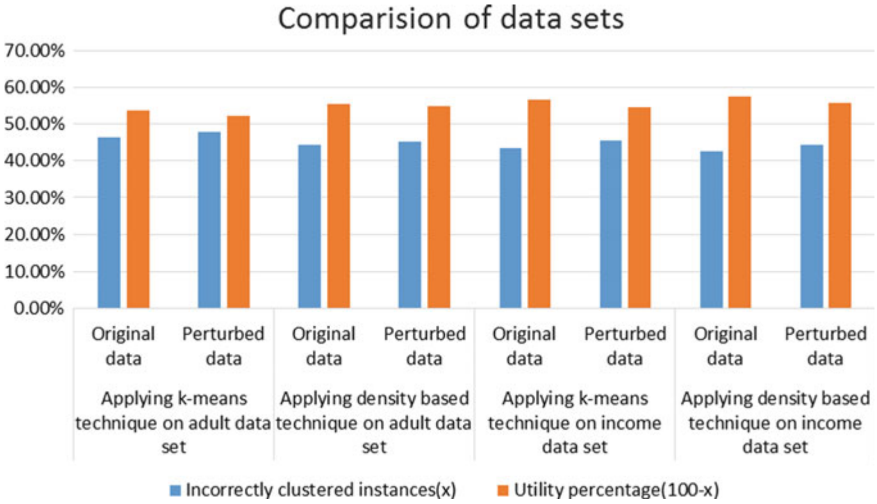


Fig. 6 Graphical representation of all data sets

to show the variation between these we have taken density based algorithm which gives better utility rates when compared to K-means.

6 Conclusion

Present article proposes performance analysis of k-means and density-based clustering technique on Heterogeneously Distorted Data. It is obtained by applying the HDD technique [16]. But there is always a mismatch between privacy and utility, when the utility is high then the privacy is low and vice versa. So in order to maintain privacy and utility, the data is heterogeneously distorted and clustered using k-means and density-based clustering techniques comparing the Incorrectly clustered instances of the original and perturbed data, it is clear that HDD can be used in PPDC to maintain a balance in privacy and the data utility.

References

1. Data Mining Curriculum. ACM SIGKDD. 30 Apr 2006. Retrieved 27 Jan 2014
2. Vaarandi, R., Pihelgas, M.: Log Cluster—a data clustering and pattern mining algorithm for event logs. In: 11th International Conference on Network and Service Management
3. Url (https://www.dataprotection.ie/docs/Offences-and-Penalties-under-the-Data-Protection-Act/g/97.htm)
4. McCue, C.: Data Mining and Predictive Analysis: Intelligence Gathering and Crime Analysis
5. Hyma, J., Prasad Reddy, P.V.G.D., Damodaram, A.: A study of correlation impact on privacy preserving data mining. Int. J. Comput. Appl. **129**(15), 22–25 (2015). (November)
6. Shah, A., Gulati, R.: Privacy preserving data mining: techniques, classification and implications—a survey
7. Malik, M.B., Ghazi, M.A., Ali, R.: Privacy preserving data mining techniques: current scenario and future prospects. In: 2012 Third International Conference on Computer and Communication Technology (ICCCT), 23–25 Nov 2012
8. Patel, D., Modi, R., Sarvakar, K.: A comparative study of clustering data mining: techniques and research challenges
9. Mythili, S.: Int. J. Comput. Sci. Mobile Comput. (IJCSMC) **3**(1) (2014) (January)
10. Mann, A.K., Navneet Kaur, R.I.M.T., Mandi Gobindgarh, P.T.U.: Software & data engineering. Global J. Comput. Sci. Technol. **13**(5) (2013) (Version 1.0)
11. Paliwal, P., Sharma, M.: Enhanced DBSCAN algorithm. Int. J. Comput. Trends Technol. (IJCTT) **4**(4) (2013) (April)
12. Meskine, F.: Int. Arab J. Inf. Technol. **9**(2) (2012) (March)
13. Shinde, S., et al.: Int. J. Comput. Sci. Commun. Netw. **4**(6), 197–202
14. Yadav, J., Sharma, M.: A review of K-mean algorithm. Int. J. Eng. Trends Technol. (IJETT) **4**(7) (2013) (July)
15. Summer School "Achievements and Applications of Contemporary Informatics, Mathematics and Physics" (AACIMP 2011) August 8–20, 2011, Kiev, Ukraine Density Based Clustering Erik Kropat University of the Bundeswehr Munich Institute for Theoretical Computer Science, Mathematics and Operations Research Neubiberg, Germany
16. Hyma, J., Prasad Reddy, P.V.G.D., Damodaram, A.: Performance analysis of heterogeneous data normalization with a new privacy metric. Int. J. Comput. Sci. Inf. Secur. (IJCSIS) **14**(6) (2016) (June)
17. https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data
18. Url (https://archive.ics.uci.edu/ml/machine-learning-databases/income/income.data)
19. Saroj, T.C., Chaudhary, T.: Study on various clustering techniques

Dimensionality Reduction Using Subset Selection Method in Framework for Hyperspectral Image Segmentation

B. Ravi Teja, J. Hari Kiran and B. Sai Chandana

Abstract This paper presents a dimensionality reduction method using subset selection for hyperspectral image segmentation framework. This framework consists of three stages—dimensionality reduction, hierarchical image fusion, and segmentation. A methodology based on subset construction is used for selecting *k* informative bands from *d* bands dataset. In this selection, similarity metrics such as Average Pixel Intensity (API), Histogram Similarity (HS), Mutual Information (MI) and Correlation Similarity (CS) are used to create *k* distinct subsets and from each subset, a single band is selected. Hierarchical fusion is used to create a single high quality image. After getting fused image, Fuzzy c-means (FCM) algorithm is used for segmentation of image. The qualitative and quantitative analysis shows that CS similarity metric in dimensionality reduction algorithm gets high-quality segmented image.

1 Introduction

Hyperspectral dataset consists of narrow remote sensing images, with each image giving information about the object on earth in a specific spectral band. The data set consists of several spectral channels raging from visible region to infrared region of electromagnetic spectral. The knowledge extracted from these images are useful in many applications such as Urban monitoring, Agriculture, Ecosystems, etc. Unsupervised algorithms for high accuracy segmentation is a new research area in hyperspectral imaging applications [1].

The framework consists of three stages in segmenting a hyperspectral data set—dimensionality reduction, fusion, and segmentation. The hyperspectral data set consists of hundreds of bands and the spectral features between adjacent bands are highly correlated, providing redundant information. The dimensionality reduction algorithms

B. Ravi Teja (✉)
GITAM Institute of Technology, GITAM, Visakhapatnam, India
e-mail: ravitejabhima@gmail.com

J. Hari Kiran · B. Sai Chandana
Shri Vishnu Engineering College for Women, Bhimavaram, India

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_52

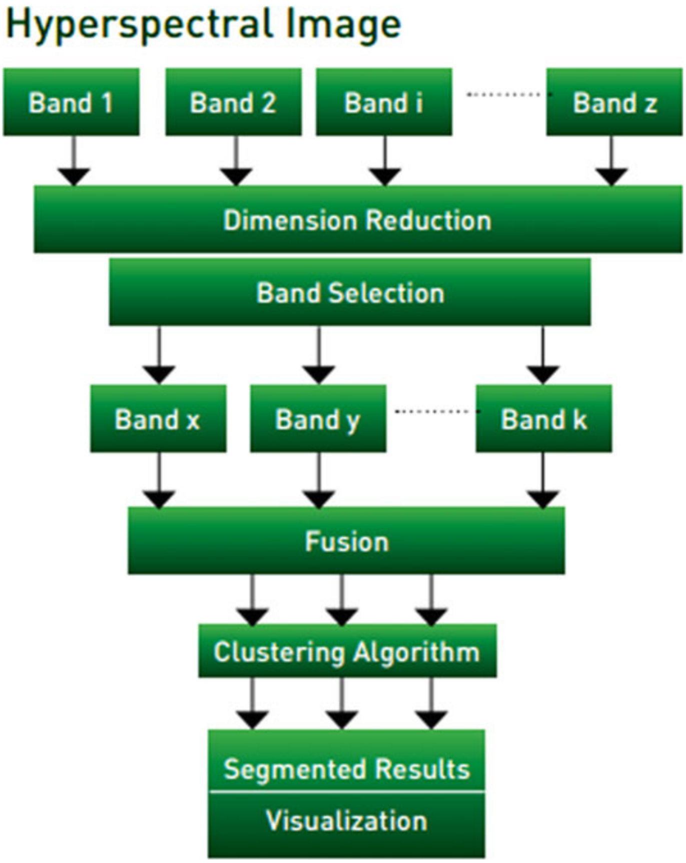


Fig. 1 Hyperspectral image segmentation framework

removes the redundant bands, thus decreasing the storage space and computational load in processing hyperspectral dataset. In this paper, a new algorithm for dimensionality reduction is presented using subset selection method for selecting K informative bands from D bands dataset. After selecting K informative bands, these K bands are fused to get an image with all the features in K bands. This fused image is segmented using FCM. The hyperspectral segmentation framework is shown in Fig. 1. This paper is organized as follows: Sect. 2 presents dimensionality reduction algorithm, Sect. 3 presents hierarchical image fusion, Sect. 4 presents FCM, Sect. 5 presents experimental results and finally conclusions.

2 Dimensionality Reduction in Hyperspectral Images

Hyperspectral data set consists of stack of images which are strongly correlated means that there is huge amount of redundant information and that data need to be removed before segmentation [2]. The dimensionality reduction can be done using transformation based methods or selection based methods. In transform based methods matrix transformations are used to project the data into lower dimension space which changes the physical meaning of spectral data. The selection-based methods directly measures the information content in each individual band. The band selection is done by choosing the bands with higher information content. Selection-based methods are better than transform-based methods, as did not change the meaning of original dataset. In transform-based methods, as the original data is transformed, some useful information required for segmentation may be distorted, changing the physical meaning of data [3]. This paper presents a new methodology of selection based dimensionality reduction of hyperspectral data.

Given hyperspectral data set with d bands. From these *d* bands, proposed methodology finds *k* bands that give us the most information for segmentation, discarding (*d*-*k*) bands. These set of *k* bands contains the least number of dimensions that most contribute to accuracy of segmentation. Objective of this work is to create subsets of bands based on similarity criteria. Most similar bands are grouped into one subset. All bands in the same subset are similar with respect to metric *M* and the bands in different subsets are dissimilar with the same metric M. In this way k subsets are created, and we select one representative band from each subset. These *k* bands from *k* subsets are used for further processing in the framework for hyperspectral image segmentation. The remaining (*d*-*k*) are discarded, thus achieving dimensionality reduction.

In-order to create subsets, three parameters are required. A metric '*M*', a reference band 'I_{ref}' first image of a subset or group, image I_i in the dataset, K numbers of bins in image histogram and Threshold 'T'.

Metric '*M*': A metric *M* is a measure that directly shows how similar two images are. The metrics used in creation of subsets are API, HS, MI and CS similarity metrics [4]. The equations of similarity metrics is given below:

$$\text{API}_i = \frac{1}{M * N} \sum_{x=1}^{M} \sum_{y=1}^{N} I(x, y)$$

$$M = |I_{\text{ref}} - I_i|$$

$$\text{MI} = \sum_{a \in A} \sum_{b \in B} p(a, b) \log \frac{p(a, b)}{p(a).p(b)}$$

$$\text{HS} = \sum_{k=1}^{K} \sqrt{I_{\text{ref}}(k).I_i(k)}$$

$$\mathrm{CS} = \frac{\sum_x \sum_y (I_{\mathrm{ref}}(x, y) - \overleftarrow{I}_{\mathrm{ref}})(I(x, y) - \overleftarrow{I})}{\sqrt{\sum_x \sum_y (I_{\mathrm{ref}}(x, y) - \overleftarrow{I}_{\mathrm{ref}})^2 (I(x, y) - \overleftarrow{I})^2}} \tag{1}$$

Reference band 'I_{ref}': It always denotes the first band of each subset. It is compared with the bands in the subset keeping similar images in the subset and discarding dissimilar images.

Threshold T: This is threshold that shows the similarity of images in a subset. Higher the threshold, smaller bands in each subset, increasing the total number of subsets.

Start with subset$_i$, ($i = 1$) containing the original data set I_j ($j = 1,\ldots, d$). Assume $I_{\mathrm{ref}} = I_1$ first band in the data set. Add each band into subset$_i$, if the similarity between reference and selected band is less than threshold T. Otherwise move to subset$_{i+1}$. Start with subset$_i = \Phi$ (empty set) and bands are added to the subset$_i$ if $M(I_{\mathrm{ref}}, I_j) \leq T$. The subset$_{i+1}$ is created according to the following formula

$$M\big(I_{\mathrm{ref}}, I_j\big) = \begin{matrix} \leq T, \ \text{assign } I_j \text{ to subset}_i \\ > T, \ \text{assign } I_j \text{ to subset}_{i+1}, \ \text{and first band in subset}_{i+1} = I_{\mathrm{ref}} \text{ for subset}_{i+1} \end{matrix} \tag{2}$$

where M is the similarity metric between two bands and j ranges from 2 to d. The same process is repeated until k subsets are created. From these k subsets, k bands are selected one from each subset.

3 Hierarchical Image Fusion Technique

The selected image bands after dimensionality reduction is fused using hierarchical image fusion technique. The procedure of hierarchical image fusion with normalized weights at each stage is presented in [5] and the flow diagram of this fusion process is shown in Fig. 2.

4 Fuzzy C-Means (FCM) Clustering

The FCM algorithm presented in [6] is used for segmentation of fused image. The assignment of pixel to the cluster and updation of centroid and membership values are done according to the given equations [7, 8].

$$\begin{aligned} &u_{ij}^m D(I_i, C_j) < u_{iq}^m D(I_i, C_q), \ q = 1, 2, \ldots, K \\ &j \neq q \end{aligned} \tag{3}$$

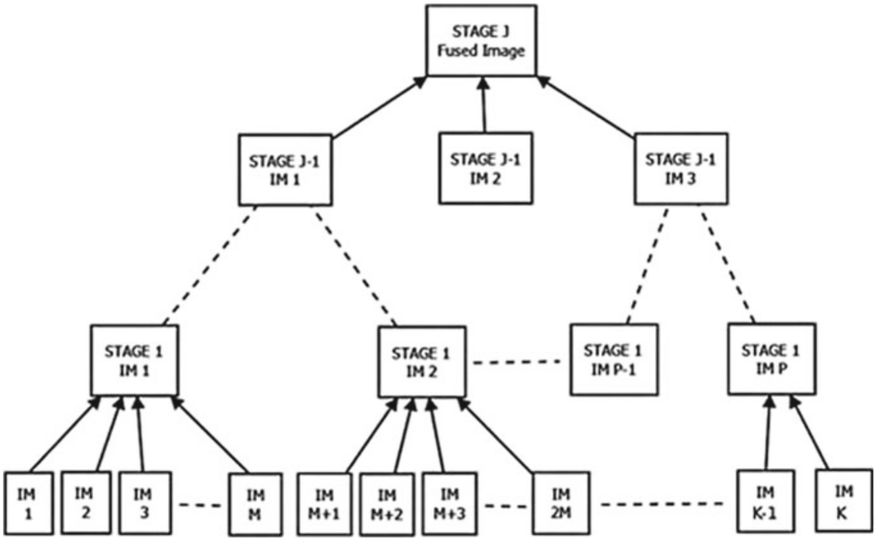


Fig. 2 Hierarchical image fusion

$$u_{ik} = \frac{1}{\sum_{j=1}^{K}\left(\frac{D(C_i,I_k)}{D(C_j,I_k)}\right)^{\frac{1}{m-1}}}, \text{ for } 1 \le i \le K \quad C_j^{\wedge} = \frac{\sum_{j=1}^{n} u_{ij}^{m} I_j}{\sum_{j=1}^{n} u_{ij}^{m}} \tag{4}$$

5 Experimental Results

The proposed methodology is tested on Pavia University data set collected from [9] with 103 spectral bands. The dimensionality reduction is done using proposed subset methodology with different similarity metrics and 40 bands selected from 103 bands. These 40 selected bands are fused using hierarchical fusion and segmented using FCM. The qualitative analysis is shown in Fig. 3 and Quantitative analysis using Mean Square Error [10, 11] with FCM clustering algorithm is shown in Table 1. With the CS similarity metric in dimensionality reduction, FCM generates lowest value of MSE. In future step, the same procedure can be repeated with different clustering algorithms with different similarity metrics used in dimensionality reduction algorithm.

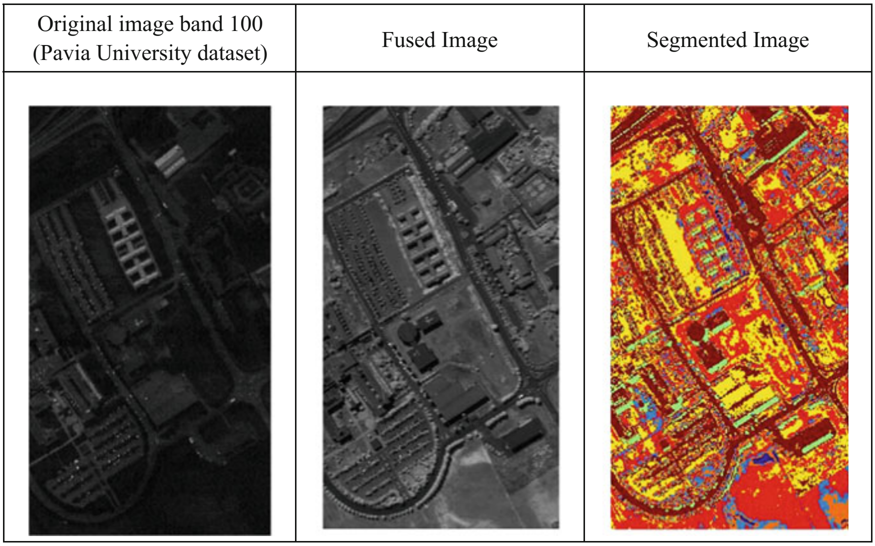


Fig. 3 Segmentation of hyperspectral image using FCM

Table 1 Quantitative analysis of image segmentation with different similarity metrics in dimensionality reduction algorithms

Similarity metric in dimensionality reduction algorithm	MSE values by FCM algorithm
API	202.8
MI	196.4
HS	192.6
CS	189.8

6 Conclusions

A framework for hyperspectral image segmentation is presented with three stages—dimensionality reduction, image fusion, and segmentation. In this paper, a dimensionality reduction method using subset selection with different similarity metrics is presented. The accuracy of any segmentation algorithm decreases if the number of spectral bands increases and existing methods for hyperspectral segmentation is carried out using limited number of bands. This proposed methodology segments the hyperspectral image by combining all the information in original dataset rather than taking specific bands. This paper also presents the performance of FCM algorithm by using different similarity metrics in dimensionality reduction algorithm.

References

1. Zhang, Z., et al.: An active learning frame work for hyperspectral image classification using hierarchical segmentation. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **3**(2), 640–654 (2016)
2. Loiencs, A., et al.: Selection of informative bands for classification of hyperspectral image: based on entropy. In: Proceedings of 15th Biennial Baltic Electronics Conference. IEEE (2016)
3. Aralan, O., et al.: A comparative analysis of classification methods for hyperspectral image: generated with conventional dimension reduction methods. Turk. J. Electr. Eng. Comput. Sci. **25**(58), 72 (2017)
4. Shah, P., et al.: Efficient hierarchical fusion using adaptive grouping technique: for visualization of hyperspectral images. In: Proceedings of ICICS. IEEE (2011)
5. Kotwal, K., et al.: Visualization of hyperspectral image: using bilateral filtering. IEEE Trans. Geosci. Remote Sens. **48**(5), 2308–2319 (2010)
6. Saicbandana, B., et al.: Image fusion in hyperspectral image classification using genetic algorithm. Indonesian J. Electr. Eng. Comput. Sci. **2**(3), 703–711 (2016)
7. Salsem, M.B. et al.: Hyperspectral image feature selection for the fuzzy c-means spatial and spectral clustering. In: Proceedings of IEEE IPAS 2016. International Image Processing Applications and System Conference (2016)
8. Saichandana, B., Srinivas, K., Kiran Kumar, R.: Dimensionality reduction and classification of hyperspectral images using genetic algorithm. Indonesian J. Electr. Eng. Comput. Sci. **3**(3), 503–511 (2016)
9. http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes
10. Harikiran, J., et al.: Multiple feature fuzzy c-mean: clustering algorithm for segmentation of microarray image. IAES Int. J. Electr. Comput. Eng. **5**(5), 1045–1053 (2015)
11. Raviteja, B., Prasad Babu, M.S., Venkata rao, X., Harikiran, J.: A new methodology of hierarchical image fusion in framework for hyperspectral image segmentation. Indonesian J. Electr. Eng. Comput. Sci. **6**(1), 58–65 (2017)

A Survey on Business Intelligence Tools for Marketing, Financial, and Transportation Services

Chavva Subba Reddy, Ravi Sankar Sangam and B. Srinivasa Rao

Abstract The main moto behind this paper is to present an elaborated survey on Business Intelligence (BI) tools for Marketing, Transportation, and Financial Services. As a survey, our objective should present much known BI tools like Tableau, Power BI, Pentaho, QlikView, and Micro Strategy Analytical Express. These tools are well known for the ease of organizational use. Comparisons are made from these tools to test their performance and try to figure out best of all. From the available data gathered from different sources, one can thoroughly conclude that Pentaho is the best in terms of data processing to services among other BI tools available in the services specified.

1 Introduction

In day-to-day with innovative technologies, enormous amount of data is exponentially generated from immense data sources. The data collected from distinct sources lead to an emergence of structured, semi-structured and unstructured types. For instance, collected data from RDBMS is structural, log files data were in semi-structured and other files such as text files, audio and video files are un-structured. Analysis and predictions of this huge amount of data is typically a difficult task. For example, it is quite difficult to launch a product into market. First one has to analyze previous similar products sales in area-wise, region-wise, and country-wise. The marketing areas of our product should be analyzed and predicted based on previous data. Nowadays, analysis of the growing data is a very difficult task and requires unpredictable period for decision making to the business executives [1].

C. S. Reddy (✉) · R. S. Sangam · B. Srinivasa Rao
Department of CSE, Vellore Institute of Technology-AP, Amaravati, India
e-mail: subba.reddy@vitap.ac.in

R. S. Sangam
e-mail: ravisankar.s@vitap.ac.in

B. Srinivasa Rao
e-mail: srinivas.battula@vitap.ac.in

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_53

Business Intelligence (BI) is the most vital for abnormal state administration of any association for break down, envision, revealing, and virtual-making arrangements including plans. BI is the most innovative approach for exploring information and introducing significant data to assist corporate administrators, business supervisors and various closure clients settled on more educated business choices.

BI as the way towards taking a lot of information and showing an effective state set of reports that pave information into the premise of business activities, empowering administration to settle on crucial day by day business choices. As an efficient service, our motive is to have a reasonable BI tools available open or closed access. The main challenging parameters in IT industry are scalability, financial commitment, usability, fault tolerance, other issues. The reliable tool should be selected in view of the above issues for organization [2, 3].

This paper explores about the suitable tools in open source or non-open source for creating the data dashboards/visualization, with low cost for developing marketing and financial organizations, and Transportation Service projects [4]. The rest of the paper is organized as follows: Sect. 2 describes Micro Strategy Analytics Express. Section 3 describes QlikView. Section 4 describes Tableau. Section 5 describes Pentaho. Section 6 describes Power BI. Finally, we draw conclusion in Sect. 7.

2 Micro Strategy Analytics Express

M. Michel, J. Saylor and Sanju Bansal founded Small-scale Strategy, Inc established in 1989. The main purpose of the company is to develop tools for the analysis of internal and external information so as to settle on better choices. The Micro Strategy is a provider of BI to Mobile and cloud-based services. Microsoft Strategy Analytics Desktop is a BI tool for business analysis developed by Micro Strategy Company. In a BI market, Micro Strategy was one of the independent, publicly trade BI Software providers. Micro Strategy products are Micro Strategy Analytics, Micro Strategy Mobile and Usher.

Micro Strategy Analytics Express permits to create interactive dashboards for visual striking to explore business data with high speed. Once making done, we can share dashboards among the team. One can view dashboards by using Analytics Express app in iPad and on windows platform users can use Internet Explorer, Firefox 13+, Chrome 20+, and on Apple platform users can use Safari 5+, Firefox 13+, and Chrome 20+.

Our objective is to exploit different types of data sources to create Dashboards for PC or Systems. The data sources such as Drop Box, URL, Google Drive, Saleforce.com, database connections (OLAP and OLTP), and CSV files, Excel Sheet, Google Spread Sheets, Oracle, SQL Server, PostgreSQL and Sybase ASE are some of the examples.

Micro Strategy Analytics permits big organizations to study large amount of data and securely obtain actionable business insights through-out an enterprise. Micro Strategy generates dashboards and reports along with conducting ad hoc analysis

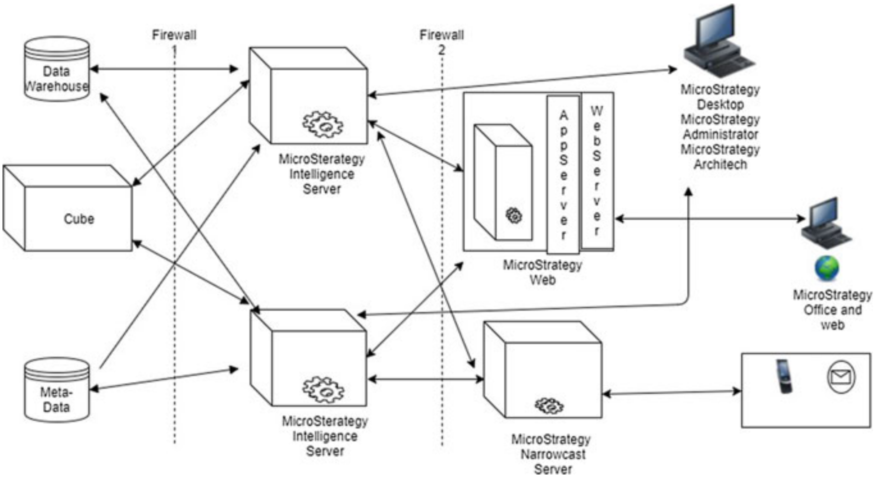


Fig. 1 Architecture of Micro Strategy Analytics Express

that allows sharing the results among mobile devices or webbing. It also has data discovery with the security, scalability and governance features of enterprise-grade BI [5].

A simple architecture of Micro Strategy Analytics Express is expressed in Fig. 1 with three stage demonstrations. The first stage depicts about data gathering from different data warehouses and databases. The second stage specifies about optimized analytical sever called Micro Strategy Intelligence used for querying, reporting and OLAP analysis. It has important functions such as sharing data including objects and metadata information protection collecting from stage one. The final stage is about creating dashboard, administration and share dashboards among the users.

Pros:

1. It is cost-effective administration tool.
2. It accesses cloud data.
3. Easy to use and maintain.
4. It accesses Big Data solutions.
5. Real-time reporting design.
6. Integrated solution for delivering different types of reports through multiple mediums like web, mobile or distribution services.
7. Advanced and predictive analytics.

Cons:

1. Version control and parallel development limitation.
2. If an employee is familiar with SQL then only one can understand how to use this Micro Strategy.
3. Merging data from different sources has been improved a lot but it is not an easy to create a query manually.

3 QlikView

Qlik is a software company in Sweden started in 1993, provides QlikView and QlikSence tools for BI. Basically, it is a PC-based tool which stands for Quality, Understanding, Interaction and Knowledge named as Quik, later it is named as Qlik.

Nowadays QlikView is leading to business discovery. These are in-memory based tools, which mean all data are loaded into the RAM. QlikView is an intuitive self-benefit representation and disclosure device to break down, translate and envision enormous information sources. The main advantages in QlikView are drags and drops feature for building dashboards, reports and visualizations. QlikView is available in three versions- Desktop, Enterprise and Cloud. QlikView is used for guided analytics and QlikSence is used for Self-Service visualization [6].

QlikView support many data formats and sources such as EXCEL, CSV files, XML files, databases, Web files, QVD (QlikView Document) files, data access from OLE DB and ODBC data sources, DIF, HTML and QVX. QlikView and QlikSence are free of cost for one personnel user. One of main advantage is current selections are saved in the form of bookmarks for later usage and users can easily access bookmarks.

In QlikView, the data visualization is in the form of charts. The different types of charts are Bar charts, Line chart, combo chart, Radar chart, Scatter chart, Grid chart, Pie chart, Funnel chart, Block chart, Gauge chart, Mekko chart, Pivot chart, Straight table. Qlik View can easily adopt immense datasets, its interface can easily lay a ladder to data sources pointing ongoing data visualization at real time. One can incorporate data from various sources and the data can rapidly be made accessible through Source specific API's. End executives can visualize data in browser by using Add-on's. Applications of QlikView are human resource administration, production control, financial systems, project administration, market analysis, customer support, stock inventories and purchasing [7].

A simple architecture of QlikView is expressed in Fig. 2. The working procedure is explained as follows. The initial step is to extract data from specific sources and integrating into QlikView. In the second stage, pre-processing is done by cleaning spurious data fields and identifying outliers. In stage three, users are facilitated with drag and drop visualization modules for dash boards. In final stage, dashboards can be shared to other customers and higher level department peoples.

Pros:

1. The main advantage of QlikView and QlikSence is in-memory based tools.
2. We can generate reports in the form of EXCEL, PDF.
3. Offers associative search capabilities.
4. Large partner and consultant base.
5. Robust mobile apps.
6. Sound dashboard technologies.
7. Easier to develop cycle than traditional and legacy BI platforms.
8. Solid integration.

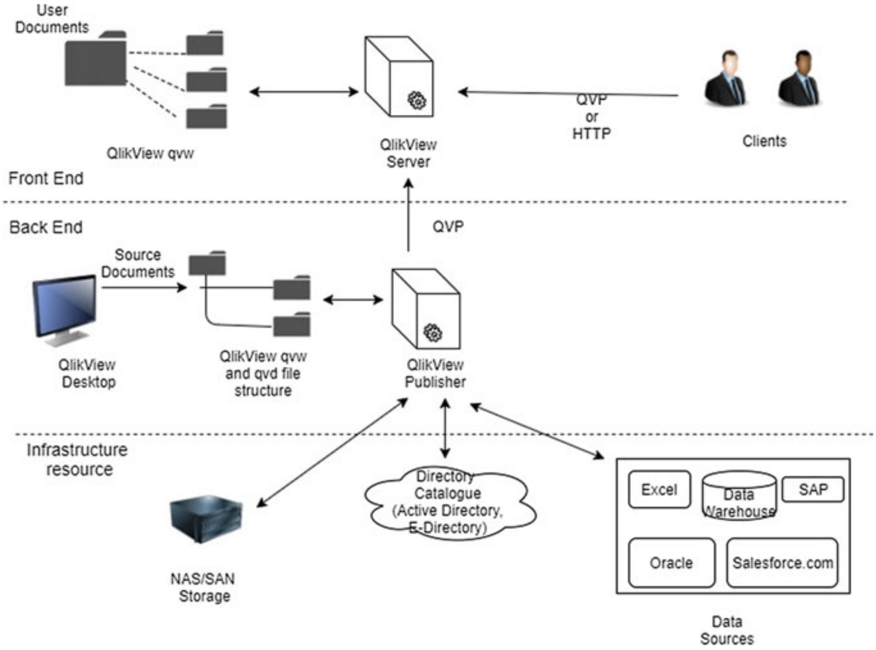


Fig. 2 Architecture of QlikView

Cons:

1. You can access large datasets depending on configuration of system.
2. Require trained developer.
3. No centralized security.
4. Challenging to embed analytics.
5. Outdated interface.

4 Tableau

Tableau is another important tool for BI software to make data analytics and visualization. Reporting can be easily done by drag-and-drop feature. Tableau is user-friendly tool and one can learn easily without a prior programming experience. Tableau was founded by Christian Chabot, Chris Stolte and Pat Hanrahan in 2003.

Tableau can perform in-memory processing like QlikView. Data analysis is done without connecting to any data sources. The amount of data analyzed in Tableau is based on availability of memory. Tableau can get connected with different data sources and access data before going to analysis. Some data sources are CSV, EXCEL,

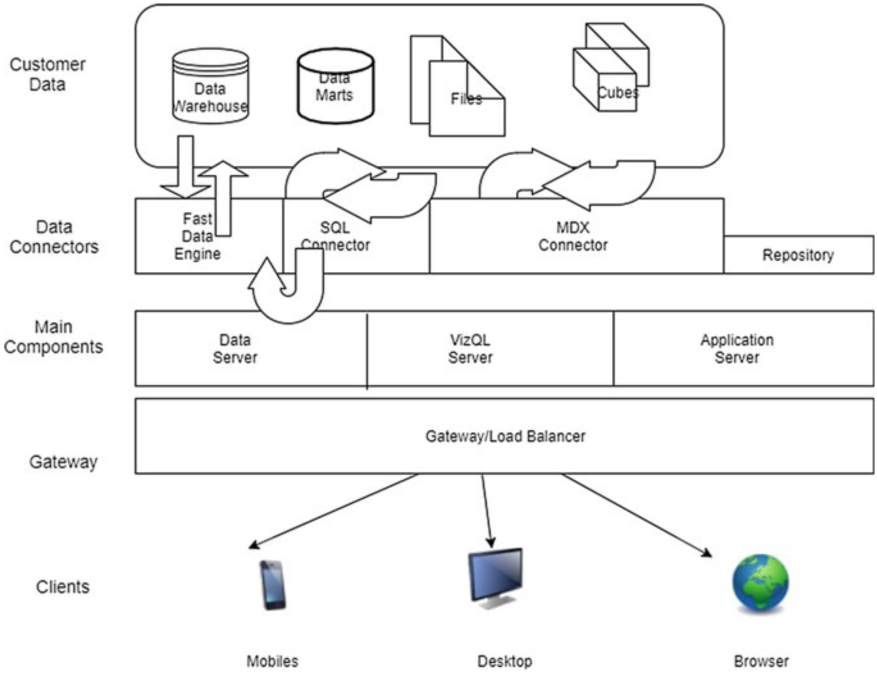


Fig. 3 Architecture of Tableau

Oracle, SQL Server, IBM DB2 and ODBC. The data can also access from cloud systems such as Windows Azure, Google Big Query, and Big Data.

Tableau can visualize data in different charts like Bar chart, Line chart, Pie chart, Cross tab, Scatter plot, Babble chart, Bullet graph, Box plot, Tree map, Bump chart, Gannt chart, Histogram, Motion charts and Waterfall charts [8].

It is not an open source product. The simple architecture of Tableau is expressed in Fig. 3. The functionality is explained as follows. The first stage is about collecting data from different sources. Second stage is about establishing connections between data sources and Application server. In stage three, the creation of dashboard with Gateway/Load balancer is done. Finally, dashboards can be shared among clients.

Pros:

1. Very user-friendly interface.
2. Easily integrated with third party.
3. Mobile support for dashboard reports.
4. User forums and customer services.
5. Low-cost development.

Cons:

1. Initial data preparation.
2. Tableau not provides all statistical features.
3. Tableau cannot replace Financial Reporting Applications.

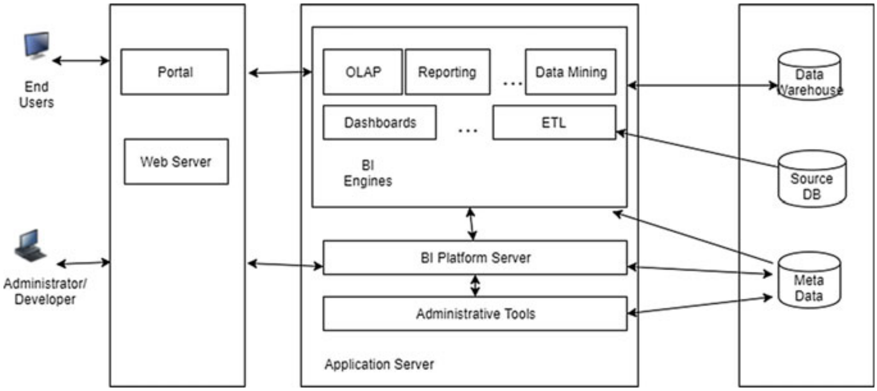


Fig. 4 Architecture of Pentaho

5 Pentaho

Pentaho is another BI software company that offers, open sources products which provide OLAP services, Information dashboards, data integration, reporting, ETL and Data mining capabilities. Pentaho was founded by five peoples in 2004 at Orland, Florida. Pentaho have an enterprise edition and community edition. Pentaho offers two types of application

- Server Application
- Desktop/Client Applications

Pentaho application products are Pentaho BA Platform, Pentaho analysis service. Pentaho have some server Plug in such as Pentaho Dashboard Designers, Pentaho analysis, Pentaho interactive Reporting, Pentaho Data Access Wizard, Pentaho desktop application products are Pentaho Data Integration (PDI), Pentaho Report Designer, Pentaho Data mining, Pentaho Metadata Editor, Pentaho Aggregate Designer, Pentaho Schema Workbench and Pentaho Design studio. Pentaho access various data sources. Pentaho by default has ODBC drivers to communicate with databases. The databases supports for Pentaho are MySQL, HypersonicSQL and Hive, etc. Pentaho generates reports in the form of HTML, XML, Excel, CSV, PDF and Text [9].

A simple architecture of Pentaho is expressed in Fig. 4. The architecture of this tool is explained as follows. In first stage, data is collected from different sources. In stage two, Extraction, Transforming and Loading are performed on the data and save into repository. In stage three, dashboards are created for reporting and analysis of data. In final stage, sharing dashboards to clients within web browsers, web portals are done.

Pros:

1. Pentaho solves complex business solutions with its cool feature.
2. Pentaho has good features including data mining and integration and ETL capabilities.
3. Pentaho provides good insight of the business.
4. User-defined code is very flexible; it is possible and very easy to plug in user-defined java code.

Cons:

1. Periodically, Pentaho Spoon (GUI) crashes, and a restart is required.

6 Power BI

Power BI provided by Microsoft for business analytical services has self-service BI capabilities that create reports and dashboards by end-users without depending on any database administration and information technology staff. Power BI provides cloud-based BI services known as Power BI administrations along with desktop-based interface called Power BI desktop. It has variety of capabilities like data warehouse, data preparation, data discovery and interactive dashboards. The main advantage of this product is the ability to load custom visualization.

Power BI data sources are SQL Server Database, SQL Server Analysis Services Database, Access Database, Oracle Database, IBM DB2 Database, Sybase Database, PostgreSQL Database, MYSQL Database, Teradata Database, SAP HANA Database, SAP Business Warehouse Server, Amazon Redshift, Impala, Snow flake, Excel, Text/CVS, Folder, JSON, XML, SharePoint Folder, Azure databases, Azure Enterprise, Sales force Objects, Sales force Reports, Google Analytics, Face book, GitHub and so on [10].

A simple architecture of Power BI is expressed in Fig. 5 and explained in three stages. In first stage, gathering data from different sources. In stage two, Business Analytics Tools like Excel and Power BI Designer to design and development of dashboards based on requirements. The final stage, sharing dashboards to clients within web browsers and mobile apps.

Pros:

1. It is affordable.
2. It is connected to a noteworthy brand.
3. Microsoft is putting assets into it.
4. It has great report perception capabilities.
5. It has broad database availability capabilities.

Cons:

1. It is best for Microsoft Excel control clients.
2. It does not deal with vast information sources well.
3. It does not take into account granularity.

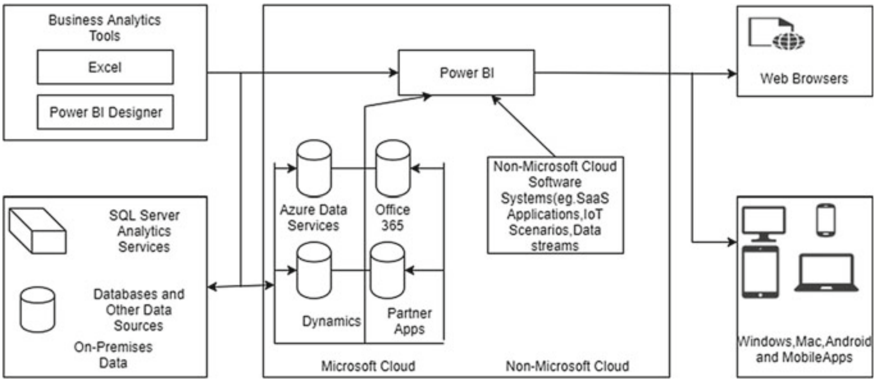

Business Analytics Tools
Excel
Power BI Designer
Power BI
Web Browsers
SQL Server Analytics Services
Databases and Other Data Sources
On-Premises Data
Azure Data Services
Office 365
Dynamics
Partner Apps
Non-Microsoft Cloud Software Systems(eg.SaaS Applications,IoT Scenarios,Data streams
Microsoft Cloud
Non-Microsoft Cloud
Windows,Mac,Android and MobileApps

Fig. 5 Architecture of Power BI

Table 1 Comparison of BI tools

Name of BI tool	Marketing services	Transportation services	Financial services
Pentaho	Good	Moderate	Good
Power BI	Good	Moderate	Moderate
Tableau	Good	Moderate	Low
Micro Strategy	Good	Low	Moderate
QlikView	Moderate	Good	Low

Finally, the performance metrics of all the BI tools are summarized in Table 1. As it is evident from Table 1 the Pentaho BI tool is more suitable for marketing and financial services.

7 Conclusion

In this paper, some of potential BI tools are discussed. These tools considered for marketing, transportation, and financial services. Pentaho tool is best for marketing, transportation, and financial services. These BI tools are very effectively, efficiently analyzing data and make decisions.

References

1. Golfarelli, M.: Open source BI platforms: functional and architectural comparison. In: International Conference on Data Warehousing and Knowledge Discovery. LCNS, vol. 5691 (2009)
2. Watson, H.J., Wixom, B.H.: The current state of business intelligence. IEEE, vol. 40, issue 9, Sept. 2007

3. Shukla, A., Dhir, S.: Tools for data visualization in business intelligence: case study using the tool Qlikview. In: Information Systems Design and Intelligent Applications, vol. 434, pp. 319–326. AIC, Feb. 2016
4. Gounder, M.S., Iyer, V.V., Al Mazyad, A.: A survey on business intelligence tools for university dashboards development. In: 3rd MEC International Conference on Big Data and Smart City (2016)
5. MicroStraregy Inc.: MicroStrategy 9: Basic Reporting Guide. MicroStrategy
6. Podeschi, R.J.: Experiential learning using QlikView business intelligence software. In: 2014 Proceedings of the Information Systems Educators Conference Baltimore, Maryland, USA. ISSN: 2167-1435
7. Garcia, M., Harmsen, B.: QlikView 11 for Developers. PACKT Publications
8. Nandeshwar, A.: Tableau Data Visualization, Cookbook
9. Tarnaveanu, D.: Pentaho business analytics: a business intelligence open source alternative. Database Syst. J. **III**(3) (2012)
10. Lachev, T., Price, E.: Applied Microsoft Power BI: Bring your Data to Life! ACM (2009)

Performance and Cost Evolution of Dynamic Increase Hadoop Workloads of Various Datacenters

N. Deshai, S. Venkataramana and G. Pardha Saradhi Varma

Abstract In the past years, Datacenters and Clusters data processing are incredibly crucial tasks. To addressing these issues, many researchers have covered up. The MapReduce is an open-source Hadoop structure expected for managing and delivering disseminated vast terabyte information on immense clusters. Its key duty is to reduce the conclusion time for large clusters of MapReduce Jobs. Hadoop Cluster has limited fixed slot design for cluster lifespan. This preset slot configuration may increase the completion time (makespan) and decrease the system resource utilization. The present open Hadoop source permits simply static slot configuration, similarly set of map slots in check to decrease slots all through the cluster lifespan. Such fixed configuration may guide to extend the completion time and the resource utilization of system will decrease. The proposed novel technique for minimizing the makespan of given set using slot-ratio among map and reduce tasks. Through utilizing the workload data of recently finished jobs it allocates slots to map and decrease tasks dynamically.

1 Introduction

The main role of a Cloud Scheduler is to allocate the resources for various jobs executing in cloud environment. Virtual machines are developed and organized in cloud to establish an environment for completion of all jobs. MapReduce is a powerful framework utilized for processing huge data applications on a group of physical machines. Presently many associations, analysts, government firms are implementing MapReduce applications going on open cloud. Implementing MapReduce application on cloud has many benefits like inception of clusters on-demand and expandability.

In general, Hadoop MapReduce open source is typically used compared with numerous MapReduce applications like Dryand, Google Map Reduce are available

N. Deshai · S. Venkataramana (✉) · G. Pardha Saradhi Varma
Department Information Technology, S.R.K.R Engineering College Affiliated to JNTUK, Bhimavaram, Andhra Pradesh, India
e-mail: vrsarella@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_54

[1–3]. However implementing a Hadoop cluster on a private group is uncommon from running on an open cloud. Open cloud empowers towards virtual group wherever assets might be provisioned or else free as expressed by the need of the application in minutes. Executing applications of MapReduce for cloud to empower client to actualize jobs of unique prerequisites excluding any sting of making physically fit preparing a cluster. In the MapReduce applications performance, the scheduling plays a key role. The standard scheduler inside Hadoop MapReduce is FIFO Scheduler [4]. But the Facebook and Yahoo uses Fair Scheduler and Capacity Scheduler respectively.

The above-stated schedulers are common cases of schedulers planned for MapReduce functionalities and are most appropriate for physical fixed clusters, that can serve the cloud frameworks through dynamic asset administration, aside from these schedulers do not think about the features influenced by virtualization concepts utilized in cloud conditions. Thus, these are the requirements of dynamic schedulers that are used to schedule MapReduce application with respect to the characteristics of the application, Virtual Machines and location of information to proficiently complete these applications in different cloud condition.

It considers that the figuring time of occupations additionally is much of the time used to effort the execution and utilize adequacy of a framework. Inside distinction, entire accomplishment spot is known to the expansion of finished up point for all employments from the time when beginning introductory occupation. It is a far-reaching makespan through lining minute (i.e., holding up time) incorporated. Utilize it to process the fulfillment in conspire as of for a solitary employment point of view through partitioning. Go for one division of development MapReduce workloads to comprise free circumstance with surprising methodologies. Proposed for dependent occupations (i.e., MapReduce work process), solitary MapReduce can simply build up just when its former dependent employments stop the computation issue to the information yield information dependence.

In contrast, for independent employments, here is an overlie division among two occupations, i.e., when the current occupation finishes its guide stage count and begins its decreased stage totaling, after that activity can start to finish its guide stage calculation (Fig. 1).

2 Related Work

Introduction enhancement implied for MapReduce Assignments is a dreadfully seen enchanting subject for scientists. Survey concerning issue to our future work Planning through Asset Designation Advancement brimming with specialists endeavor on improvement exertion for MapReduce employments, and remunerated enthusiasm on count setting up and asset distribution matter of the comparable. Numerous authors measured errand requesting enhancement utilized for MapReduce workloads [5–8]. The displaying of the MapReduce in two arrange half and half stream is clarified in [9].

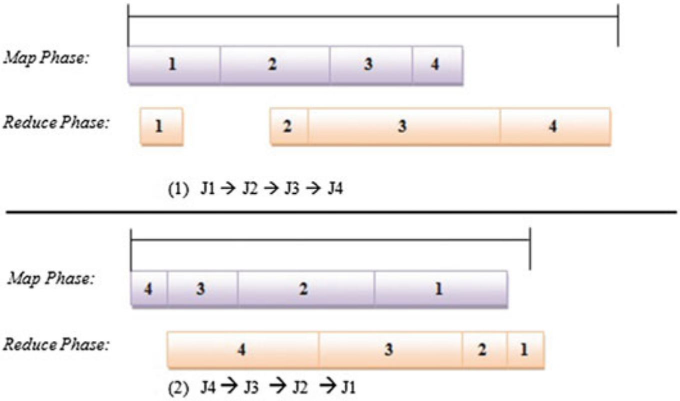


Fig. 1 MapReduce execution flow used for unlike job orders

The execution point for mapping and decreasing the employments for each activity need to perceive some time recently, yet this event is not actualized in the applications. Additionally this procedure is not measured for the significant occupations and appropriate just for the autonomous employments. If there should arise an occurrence of such strategy it is MapReduce work process. Contrasted with this marvel, our proposed DHSA is fit for the whole sorts of employments.

Starfish [10] structure can change the Hadoop design mechanically for the MapReduce occupations. By utilizing testing method and cost-based model we can abuse the use of Hadoop bunch. Yet at the same time could recoup the show of this procedure by augmenting the abuse the guide and by lessening spaces. Polo et al. [11] expected a strategy for MapReduce multi-work workloads based by asset ready gauge system.

YARN [12] clarifies the inadequacy inconvenience for the Hadoop MRv1 amid the impression of asset administration. As a substitution for opening, it oversees assets into compartments. The guide in addition to lessen activities is complete on each holder. Theoretical Execution advancement in MapReduce needs assignment game plan approach for managing in the midst of tribulations since straggler issue for a specific employment, which contains LATE [13], BASE [14], Mantri [15], MCP [16] approximate execution be a fundamental errand setting up technique.

Guo et al. [14] proposes an Advantage Mindful Theoretical Execution (BASE) calculation which assesses the conceivable advantages of the inexact errands and the unnecessary runs are dispensed with. This BASE calculation of the assessing and disposal can show signs of improvement the introduction for not on time. The theoretical execution technique extends its attention generally on sparing bunch figuring asset, is given by Mantri [15]. MCP is a most recent theoretical execution calculation arranged by the Chen et al. [16] foreseen for fitting the trouble that was irritating the introduction of the past temporary finishing strategies. Here and proposed theoretical Execution Advancement technique that adjusts the tradeoffs among lone errands and accumulation of employments. Information Region Streamlining numerous past work on adapting the introduction and adequacy of the use of group have shown to be basic undertaking. There are two range approaches utilized for Guide Lessen, diminish side in addition to outline.

In delineate information region approach, Guide work estimation is roused by the enter information (for case in [17, 18]). In [17, 19], the hypothesis of misfortune Scheduler is second hand to mind master the data locale assignment. The occupations of MapReduce to be ordered into three assortments, delineate grave at that point outline decrease input profound likewise then after the lesser input substantial through Purlieus [20].

3 Issue Definition

To misuse the opening use for MapReduce and strength the show tradeoff among a lone activity and an arrangement of occupations by reasonable setting up and socializing the show of MapReduce bunch in Hadoop. The thought is to influence use of the spaces inside MapReduce to bunch. The space uses leftovers a difficult undertaking because of uniformity likewise asset necessities. It is to be reasonable when the sum total of what pools have been owed with the comparable total of benefits. The assets necessities among the guide space and reduce opening are ordinarily divergent. This is since the guide assignment and decrease errand are frequently exhibit at long last unique completing examples.

4 Proposed Framework

See Fig. 2.

5 MapReduce Programming Model

This model was anticipated in 2004 by the Google, which is utilized as a part of preparing with creating massive informational collection execution. This system tackle set of issues, similar to information dispersion, work planning, mistake resistance, machine-to-machine correspondence, etc. Mapper Guide work requires the client to hold the incentive to a couple off of key in esteem and make a rest of halfway key and also esteem sets <key, value> comprises of two sections, esteem implied for the information associated to the activity, key implied for the “gathering number “ of the esteem. MapReduce join the middle of the road esteems with enter at that point dispatch them in the track of diminish work. Guide calculation system is depicted as following:

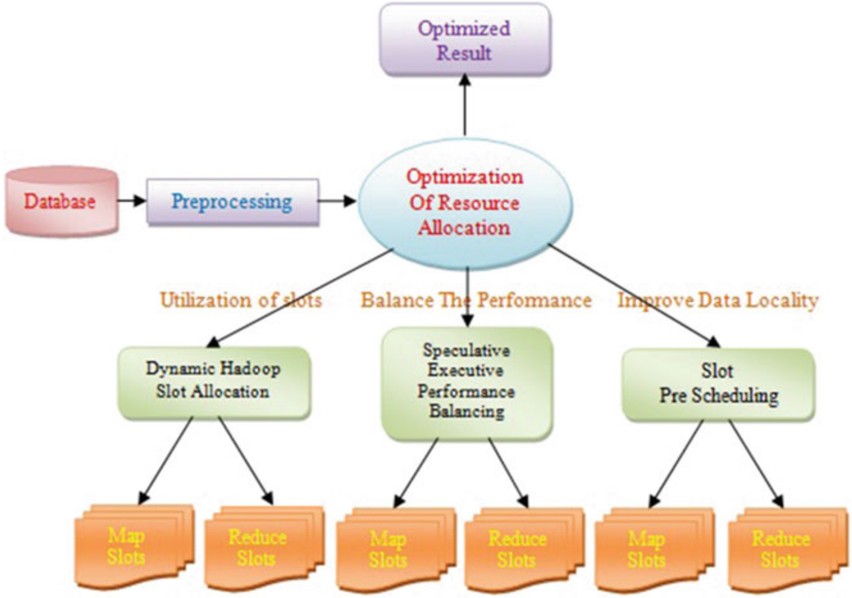


Fig. 2 System architecture

Map Algorithm Method:

Step 1: Hadoop and MapReduce structure produce a guide assignment in each Information Split, and together Info Split is cause by the Info mastermind of employment. Each <Key,Value> compares for a guide errand.
Step 2: Perform Guide undertaking, process the info <key, value> to structure a most recent <key, value>. This methodology is called “partition into gatherings”. That is, influence the related esteems to coordinate to the comparative watchwords. Yield key esteem matches that do not require the indistinguishable kind of the information key esteem sets. A known info esteem combine might be mapped into 0 or other yield sets.
Step 3: Mapper’s yield is sort to be because of all Reducer. The whole measure of pieces and the amount of occupation diminish assignments is the alike. Clients can execute divided interface to deal with which key is apportion to which Reducer. Reducer diminish work is additionally managed through client, which feels the middle of the road key sets with the esteem set aligned with the transitional key esteem. Decrease work mergers these qualities, to get a smaller than normal arrangement of qualities, the procedure is called “blend”. However this is not simple gathering. There are troublesome operations in the strategy. Reducer makes a gathering of middle person esteems set that is related through the comparable key littler. In MapReduce structure, the developers do not require to worry on the certainties of information articulations, so <key, value> is the correspondence interface for software engineer

in MapReduce show. The <key, value> ready to be viewed as a "note", key is the letter's migration address; esteem is the letter's substance.

With the comparative address composing soul be conveyed to the unaltered place. Developers solitary require to situate up effectively <key, value>, MapReduce structure can over and again and accurately bunch the qualities by coordinating key together. Reducer calculation technique is communicated as takes after:

Reducer Algorithm Method:

Step 1: Fight, contribution of Reducer is the yield with arranged Mapper. In this stage, Guide Diminish will distribute related wedge for each Reducer.
Step 2: Sort, in this progression, the contribution of reducer is swarmed by the key uncommon mapper may have. The two phases of fight and assortment are composed.
Step 3: Auxiliary Sort, if the key gathering guideline in the halfway system is different from its law by decrease.

6 K-Means Over MapReduce

In this part, at first clarifies basic or Direct K-Means grouping calculation. Next, I will clarify circulated K-implies calculation for grouping together above Hadoop structure. K-implies calculation for bunching gets closer beneath parcel strategy for grouping wherever one-level (un-settled) conceiving the information positions is framed. In the event that K is the favored numeral of groups, all things are considered segment approaches normally find each K bunches on one time. K-implies is bolstering on the arrangement that a center position can put for a bunch. Inside demanding, for K-implies I utilize the prospect of a centroid, which be the mean or else center purpose of a gathering of focuses. Basically k-implies calculation is demonstrated as follows.

K-means Algorithm

Input: K number of cluster, D Top N documents

Output: K clusters of documents Algorithm

Step 1: Generate K centroid C1, C2, …Ck by at random selecting K documents from D replicate until here no modify in cluster among two succeeding iterations.

Step 2: For every paper di in D for j = 1 to K Sim(Cj, di) = Find cosine connection between di and Cj end for allocate di to cluster j for which Sim(Cj, di) is utmost end for.

Step 3: Update centroid in every cluster end loop Step4: end K-Means (Table 1; Fig. 3).

Slot Pre-scheduling

It enhances the opening utilization viability alongside introduction through acculturating the data region planned for delineate however recognition the fairness.

Table 1 Implementation of k-means algorithm (using $K = 2$)

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

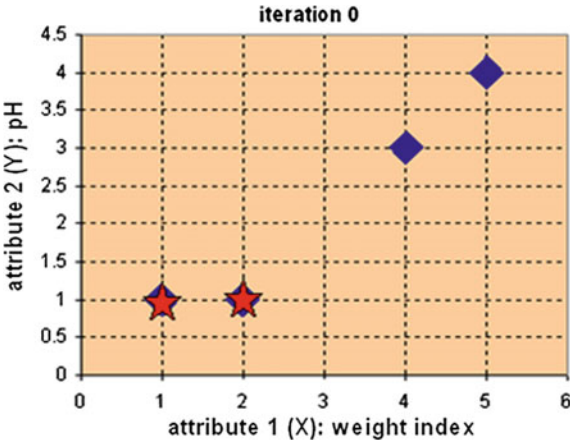


Fig. 3 Initial values of centroids

Step 1: Calculate weight aspect mapSlotsLoadFactor = awaiting map works + in succession map works commencing the entire jobs alienated by way of the group map slit capability.
Step 2: Calculate existing utmost quantity of serviceable map slots = quantity of map slots within a worktracker * minmapSlotsLoadFactor, 1.
Step 3: Calculate existing acceptable inactive map (or reduce) slots intended for a worktracker = atmost quantity of serviceable map slots—existing quantity of used map (or decrease) slots.

7 Dynamic Hadoop Slot Allocation

To boost space use while keeping up the reasonableness, when here are anticipating assignments (e.g., outline or lessen undertakings). Split verifiable explanation of MapReduce that the guide errands can simply booked guide openings and decrease occupations can essentially keep running on lessen spaces [21, 22]. In our anticipated framework change it guides with diminishing assignments that might keep running on also outline decrease openings.

Here are 4 cases, trust, NM = total of Guide errands, NR = entire entirety of Lessen undertakings, SM = entire aggregate of guide spaces, SR = whole measure of diminish openings.

Case 1: NM = SM and NR = SR The map jobs which are running on map slots and decrease jobs are run on shrink slots, There is rejection make utilize of map with reduce slots.
Case 2: NM > SM and NR < SR I assure decrease works is to decrease slots initial along with afterward utilize individuals inactive decrease slots intended for in succession map works.
Case 3: NM < SM and NR > SR I am planed individuals idle map slots intended for management reduce tasks.
Case 4: NM > SM and NR > SR The structure present within entirely active condition.

Conditional execution introduction relating it perceive the opening store inadequacy emergency expected for a Hadoop bunch, impact through speculative works. It works going on pinnacle of the Hadoop provisional scheduler toward balance the execution tradeoff among a sole work in check with gathering of employments. Space Pre-Booking enhances the opening utilization adequacy and execution through acculturating the data position proposed for outline while recognition the balance. Conditional works be trying for persuaded belonging including system, delineate alongside diminish works.

For augmenting the execution should finish the anticipating obligations first before considering the theoretical errands. At the point when hub is having inactive guide space, at that point I ought to consider anticipating map work initially and afterward consider theoretical guide errand. For a sit out of gear delineate, I initially check employments $J1$, $J2$ … Jn for outline. For each activity check the general amount of pending guide in addition to lessen errands by considering all occupations from Ji and Jj.

Where $n = 1, 2, 3, 4\ldots$

$j = i + \text{max Num Of Jobs Checked For Pending Tasks} - 1$.

Here I checked each occupation Ji by considering three conditions: (1) Add up to pending guide assignments are more noteworthy than zero. (2) No fizzled pending guide in addition to decrease undertakings for work Ji. (3) Add up to pending lessen errands is more noteworthy than zero.

8 Mathematical Model

The numerical wording of future structure is clarified as appeared beneath:

Give S a chance to be the future framework S = {I, O, F, Fs, Fl,?}

Distinguish the sources of info I: I = {T1, M, T2, R, U, E}

Where, T1 = Pending Guide Errands. M = Sit without moving Guide Spaces. T2 = Pending Decrease Errands. R = Sit out of gear Diminish Openings. U = Used Spaces. E = Purge Spaces.

Recognize set of Capacity: Let F is set of Capacities. F = {F1, F2, F3}

Where, F1 = Confirm Data. F2 = dynamic opening designation. F3 = harmony the giving of Occupation.

Distinguish the Yields: Let O be the arrangement of yields. O = {O1, O2}

Where, O1 = Spaces Dispensed Effectively, O2 = effectively harmony the look of Employment.

9 Experiments and Results

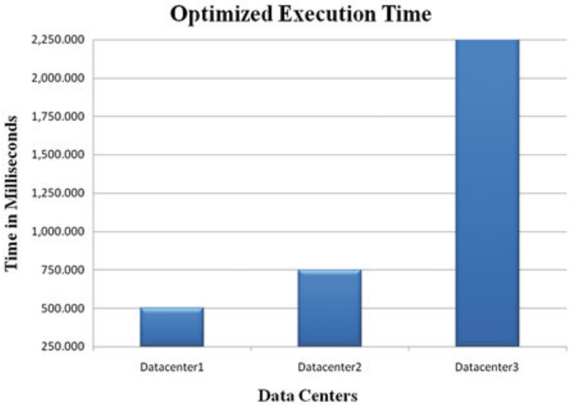
Here the standard undertaking execution time for outline/errands of all occupations as contribution to upgrade the execution of a workload. Practically speaking, the execution event of guide/lessen errands can be changed here and there the normal esteem, contingent upon particular applications. In this segment, the contact of such a minor departure from the general creation of a workload for our activity requesting calculations.

In our trial beneath except that it takes after the uniform circulation for the variety level of guide/lessen undertaking execution example with respect to the relating standard assignment execution time. The tried workload of 100 occupations and the engineered workload of 100 employments as cases to ponder the impact of various undertaking finishing time minor departure from influence traverse and entire end to time. Here produce the particular execution time for each guide/decrease undertaking with the way that takes after the uniform likelihood dissemination for the variety level of execution time subject to the given variety space.

For instance, Point 12 at the x-pivot in figure implies that the execution for each guide/decrease undertaking vacillates up or down haphazardly inside 12% in respect to the normal occupation finishing time with indistinguishable shot sharing. Subsequently, Point 0 at the x-pivot speaks to the execution comes about under the spotless standard assignment execution time. In addition standardize the makespan (or whole end time) with makespan speedup (or aggregate culmination time speedup) through isolating the makespan (or aggregate consummation time) of initially unoptimized work arrange by the one of streamlined one with our activity requesting calculations. Especially, it is worth specify that the activity composes enhancement performed is as yet base through standard errand finishing time (Fig. 4).

The aftereffects of figure demonstrate that the little increment for makespan (and add up to culmination time) for together instances of un advanced employment arrange and streamlined one with our proposed work requesting calculations under various degrees of variety in outline/errand execution time, though the relative execution change (i.e., makespan speedup, entire conclusion occasion speedup) is fine or shockingly better under this case. It shows that may land the normal position finishing time as contribution for work requesting calculations under the instance of mistake in assessing map/decrease assignment time, accepting that the variety level of undertaking execution point takes after uniform likelihood conveyance.

Fig. 4 Optimized execution time in different datacenters



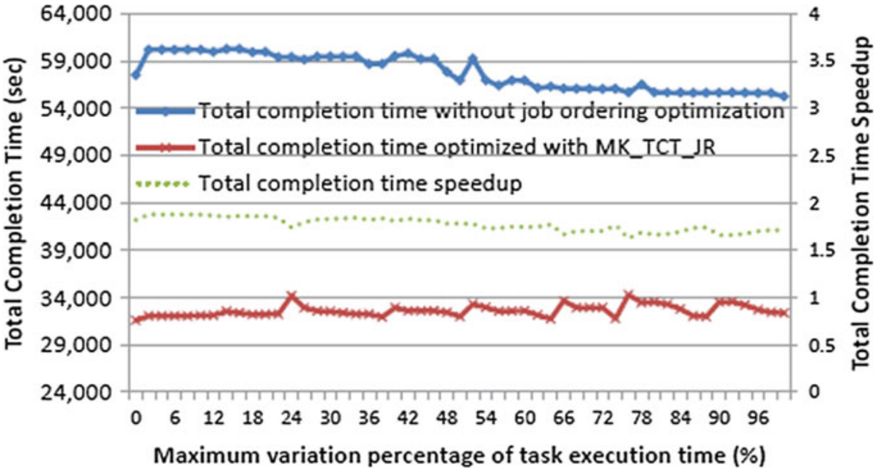


Fig. 5 Overall end instance for the synthetic workload of 100 jobs

Figure 5 demonstrates the general important minute in time toward the finish of undertaking for proposed framework interestingly with the current framework which is Greatest Cost Execution. It additionally demonstrates the particular time required for the whole three systems dynamic Hadoop opening allocation (DHSA), conditional usage introduction coordinating (SEPB) fine opening Pre-Planning.

10 Conclusion

Proposal of the projected arrangement in the route of progress is the presentation to MapReduce workloads. It is well throughtout three techniques: active Hadoop period allotment, tentative implementation presentation matching, and period Pre-Scheduling. Active Hadoop period allotment use allotment of record toward make top utilize of the period consumption, it reduces the assignment vigorously. It does not require all past information or else whichever supposition along with it for existance sprint resting on whichever type of MapReduce jobs. Speculative implementation presentation matching identifies the period incompetence crisis. It manages the balance between single and collection of jobs vigorously. Period Pre-Scheduling is to be used toward improvement of the competence in period use by way of maximizing information position. I could select the use by way of totaling over idea within conventional arrangement. In the future, it can sketch toward implement of above-mentioned concept by cloud environment.

References

1. Apache Hadoop. http://hadoop.apache.org
2. Hadoop Appropriated Document Framework. http://hadoop.apache.org/hdfs
3. J. Senior member, Ghemawat, S.: Mapreduce: rearranged information preparing on huge groups. OSDI '04, p. 137150 (2004)
4. Moseley, B., Dasgupta, A., Kumar, R., Sarl, T.: On planning in outline and stream shops. In: SPAA'11, pp. 289–298 (2011)
5. Verma, A., Cherkasova, L., Campbell, R.H.: Arranging a Troupe of MapReduce employments for limiting their Makespan. In: IEEE Exchange on Reliance and Secure Processing (2013)
6. Verma, A., Cherkasova, L., Campbell, R.: Two sides of a coin: upgrading the timetable of MapReduce occupations to limit their Makespan and enhance group execution. In: IEEE MASCOTS, pp. 11–18 (2012)
7. Tang, S.J., Lee, B.S., He, B.S.: MR order: adaptable occupation requesting improvement for online MapReduce workloads. In: Euro-Par'13, pp. 291–304 (2013)
8. Tang, S.J., Lee, B.S., Fan, R., He, B.S.: Dynamic occupation requesting and opening designs for MapReduce workloads, CORR (Specialized Report) (2013)
9. Oguz, C., Ercan, M.F.: Planning multiprocessor assignments in a two stage stream shop condition. In: Procedures of the 21st Worldwide Meeting on PCs and Modern Building, pp. 269–272 (1997)
10. Herodotou, H., Lim, H., Luo, G., Borisov, N., Dong, L., Cetin, F.B., Babu, S.: Starfish: a self-tuning framework for huge information investigation. In: CIDR11, pp. 261C272 (2011)
11. Polo, J., Castillo, C., Carrera, D., et al.: Asset mindful versatile planning for MapReduce bunches. In: Middleware'11, pp. 187–207 (2011)
12. Apache Hadoop Next Gen MapReduce (YARN). http://hadoop.apache.org/docs/current/hadoop yarn/hadoopyarn-site/YARN.html
13. Zaharia, M., Konwinski, A., Joseph, A.D., Katz, R., Stoica, I.: Improving MapReduce execution in heterogeneous environments. In: OSDI'08, pp. 29–42 (2008)
14. Guo, Z.H., Fox, G., Zhou, M., Ruan, Y.: Improving resource utilization in MapReduce. IEEE Cluster **12**, 402–410 (2012)
15. Ananthanarayanan, G., Kandula, S., Greenberg, A., Stoica, I., Lu, Y., Saha, B., Harris, E.: Getting control over the exceptions in delineate utilizing mantri. In: OSDI10, pp. 1–16 (2010)
16. Chen, Q., Liu, C., Xiao, Z.: Enhancing MapReduce performance using savvy theoretical execution system. In: IEEE Exchanges on PC (2013)
17. Zaharia, M., Borthakur, D., Sarma, J., Elmeleegy, K., Schenker, S., Stoica, I.: Defer planning: a basic system for achieving locality and decency in bunch booking. In: EuroSys10, pp. 265–278 (2010)

18. Guo, Z.H., Fox, G., Zhou, M.: Examination of information locality in MapReduce. In: IEEE/ACM CCGrid12, pp. 419–426 (2012)
19. Tan, J., Meng, S.C., Meng, X.Q., Zhang, L.: Enhancing reduce task data region for consecutive MapReduce occupations. In: IEEE Infocom13, pp. 1627–1635 (2013)
20. Palanisamy, B., Singh, A., Liu, L., Jain, B.: Purlieus: Locality aware resource portion for MapReduce in a cloud. In: SC11, pp. 1–11 (2011)
21. Guo, Z.H., Fox, G., Zhou, M.: Investigation of information region and reasonableness in MapReduce. In: MapReduce12, pp. 25–32 (2012)
22. Hammoud, M., Sakr, M.F.: Region mindful diminish task scheduling for MapReduce. In: IEEE CLOUDCOM11, pp. 570–576 (2011)

Selection of Commercially Viable Areas for Taxi Drivers Using Big Data

Kavya Devabhakthuni, Bhavya Munukurthi and Sireesha Rodda

Abstract Surface transportation in urban cities is inevitable to move from one place to another place for carrying out regular activities. Taxis are assumed as one of the essential parts for transportation in New York. This paper focuses on the selection of the top profitable areas using New York City (NYC) taxi trips dataset. The data used in the current work is captured from the NYC taxi and analyzed using Hadoop Big Data to find the profitable locations for the taxi driver, so that they can increase their income by waiting in most profitable locations.

1 Introduction

Transportation has contributed a lot to the development of the society in the economic and social fields and helped in uplifting their conditions. Today, transportation has become a basic necessity for many people in many urban areas. There are various modes of transportation which are available everywhere. In large cities like United States, the taxi mode of transportation plays an important role. These taxis are used as the best substitute for the general public use of transportation to get their necessities. The taxi cabs of New York City come in two varieties, viz., yellow and green. The yellow cabs will be able to pick up passengers anywhere and the green taxis are allowed to pick up passengers in upper Manhattan, the Bronx, Brooklyn, Queens, and Staten Island. Taxicabs are operated by the private companies and most of them are licensed by the New York City taxi and limousine commission (TLC).

Analyzing the dataset of taxis such as pick-ups, distance, time efficiency, and cost are important in order to provide a good taxi service. The New York cab data contains a very huge amount of data and hence this data is considered as big data.

The main motive of this paper is to increase the profits of the taxi drivers by presenting the busiest and most profitable locations for them. This analysis is on the

K. Devabhakthuni (✉) · B. Munukurthi · S. Rodda
Department of Computer Science and Engineering, GITAM Institute of Technology, GITAM, Visakhapatnam, India
e-mail: Srikanthd1997@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_55

pick-up and drop-off locations as pick-up location is a combination of both latitude and longitude and same as drop-off locations by considering various scenarios. Generally, the basic scenario for this analysis is that the area is considered as the most profitable area, if it generates more revenue and consists of less number of empty taxis. This will help the customers by reducing their waiting time for the taxis in the busiest times.

With the increase in urbanization and more sophisticated taxi services, the volume of data captured by the New York City (NYC) dataset became too difficult to analyze using relational database management system. In order to overcome this problem, Hadoop services were utilized to effectively analyze such an extremely large dataset. Hadoop is used to store and analyze a large data of yellow and green taxis in the city of New York within a short period of time and displays the most profitable locations in the web application using Google Maps(R) to aid the drivers to navigate the nearby profitable locations.

The remainder of the paper is organized as follows: Sect. 2 describes about the literature of the review. Section 3 gives detailed information of the methodology and technologies used in order to find the best profitable locations. Section 4 shows the results and discussions. Section 5 depicts the conclusions and future scope of work.

2 Related Work

Few authors have already studied the problem of Big Data technologies and the solutions to provide better solutions to real-world problems.

A. Shvachko et al. [1] have described the architecture of Hadoop Distributed file system and report on experience using HDFS to manage 25 petabytes of enterprise data at search engine Yahoo. They acquainted that, when working with large datasets, copying data into and out of a HDFS cluster is daunting. HDFS provides a tool called DistCp for large inter/intra-cluster parallel copying.
B. Patel and Chandan [2] have described the prevailing focus on the dataset of New York City taxi trips and fare. They analyzed NYC Taxi dataset using MapReduce Framework in order to find the superlative practice to follow and to derive the output from the data which would eventually aid the people the people to commute via taxis. The quantification of Drop-offs and Pick-ups were performed which was ultimately used to obtain Driver Fare using Hive and Pig technologies.
C. Sun and McIntosh [3] also performed analysis to understand the traffic and travel patterns, including geographical and temporal components to the NYC taxi data. A general architecture has been proposed by the authors. The framework allowed users to select the type of service required and analyze GPS-related data to provide better QoS to the users in identifying the best location to find taxis.
D. Aslam et al. [4] have proposed a design which uses a roving sensor network from taxi probes in order to accurately infer traffic patterns. They used static and dynamic sensors where static sensors are placed in fixed locations in order to

collect information about traffic and dynamic sensors are attached to the vehicles to collect the information about any individual vehicles.

3 Methodology

3.1 Technologies Used

Big data basically refers to a huge amount of data that can neither be stored nor processed using the traditional methods within a given period of time. Relational database management systems and desktop systems cannot handle such big data.

This section fetches the information of the technologies used in analyzed dataset. The following information describes about some of the technologies used in this paper such as Hadoop, Hive, Hdfs, and Sqoop.

A. Hadoop: It is open-source software and it is distributed under Apache. It is software which is used for storing and processing of large datasets and it supports running applications on big data. It is a clustered system which can store and process many files with large data. Hadoop supports a distributed model which contains a single master and multiple slaves.
B. Hive: The Apache Hive is a data warehouse infrastructure which is built on top of Hadoop. It was initially developed by Facebook. Hive is used for processing structured data and it supports analysis of large datasets by internally converting the hive queries to MapReduce jobs for data processing.
C. HDFS: Hadoop distributed file system is the storage unit of the Hadoop derived from Google file system. It is designed to provide a fault-tolerant file system and it works on commodity hardware.
D. Sqoop: Sqoop is a part of Hadoop ecosystem and it is built on top of the MapReduce. The sqoop acts as an interface between the relational database management system (RDBMS) and Hadoop distributed file system (HDFS) for importing and exporting data.

3.2 Extraction of Source Data

The source data consists of yellow taxi trip records which include fields capturing pick-up and drop-off dates/times, pick-up, and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts as shown in Fig. 1. The datasets were collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP). Generally, data can be of two types', viz., historic data or real-time streaming data. The source data considered here is historic data. The data is first loaded into Hadoop Distributed File System

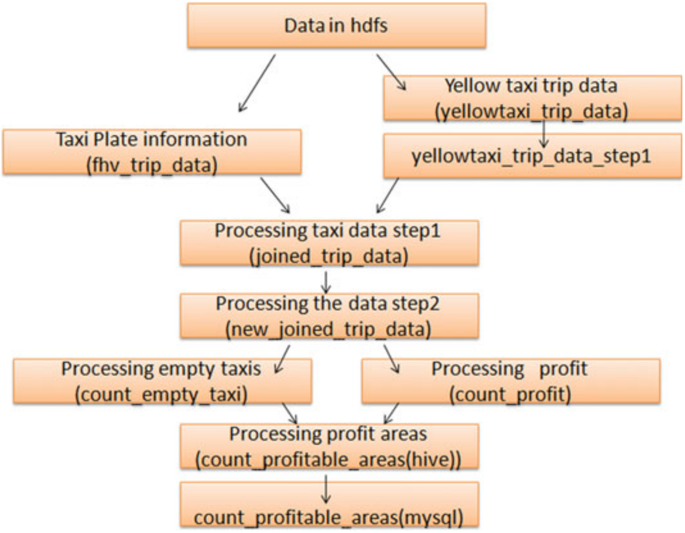

Data in hdfs
Yellow taxi trip data (yellowtaxi_trip_data)
Taxi Plate information (fhv_trip_data)
yellowtaxi_trip_data_step1
Processing taxi data step1 (joined_trip_data)
Processing the data step2 (new_joined_trip_data)
Processing empty taxis (count_empty_taxi)
Processing profit (count_profit)
Processing profit areas (count_profitable_areas(hive))
count_profitable_areas(mysql)

Fig. 1 Flowchart

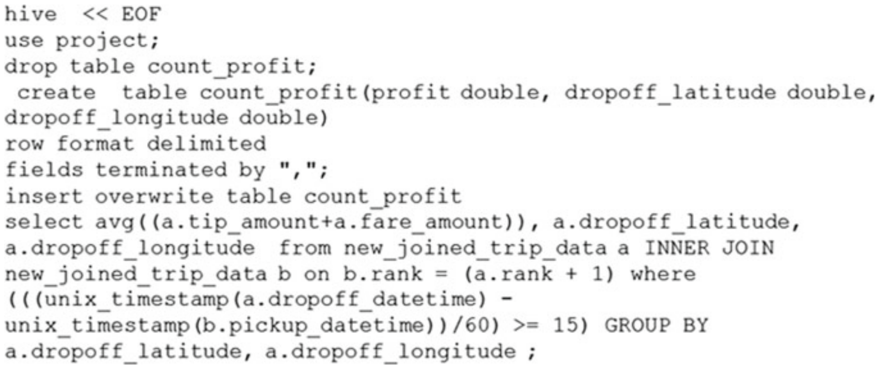
using Flume which is a real timing streaming engine. Hadoop Distributed File System (HDFS) is used to store huge amounts of structured, semi-structured or unstructured data. The data used is structured data of two huge files FHV (For Hire Vehicle) trip data which includes taxi plate information and Yellow taxi trip data. The survey is conducted by two companies creative Mobile Technologies and VeriFone Inc.

3.3 Geographic Analysis

Geographic Analysis consists of analyzing the various geographic conditions like temperature, weather, and location information. GPS tracking is used in almost all New York City taxis, which are regulated by the TLC. From the past few years, the TLC has provided access to its trip database containing millions of taxi trips with the spatial information acquired by GPS.

3.4 Hive Processing

Hive external tables are developed on the data in HDFS such as fhv_trip_data on FHV data and yellow_taxi_trip_data on yellow taxi data as depicted in Fig. 1. The records obtained from Creative Mobile Technologies were considered and a separate hive external table is created on that data which is yellow_taxi_data_step1. By applying the



```
hive  << EOF
use project;
drop table count_profit;
 create  table count_profit(profit double, dropoff_latitude double,
dropoff_longitude double)
row format delimited
fields terminated by ",";
insert overwrite table count_profit
select avg((a.tip_amount+a.fare_amount)), a.dropoff_latitude,
a.dropoff_longitude  from new_joined_trip_data a INNER JOIN
new_joined_trip_data b on b.rank = (a.rank + 1) where
(((unix_timestamp(a.dropoff_datetime) -
unix_timestamp(b.pickup_datetime))/60) >= 15) GROUP BY
a.dropoff_latitude, a.dropoff_longitude ;
```

Fig. 2 Pseudo code for identifying profit generated in the locations

inner join query on both the hive tables, a new hive table is formed joined_trip_data for the further analysis.

Based on the drop-off date, time, and licence number, the records were ranked and ordered by the newly formed field rank. Another hive table was built on this processed data new_joined_trip_data. The taxi type in data provided was the taxi which was used by different people who board at different pick-up locations and they all get dropped at the same drop-off location. The scenario considered here to find profitable locations with less number of empty taxis and the more revenue is generated from that area.

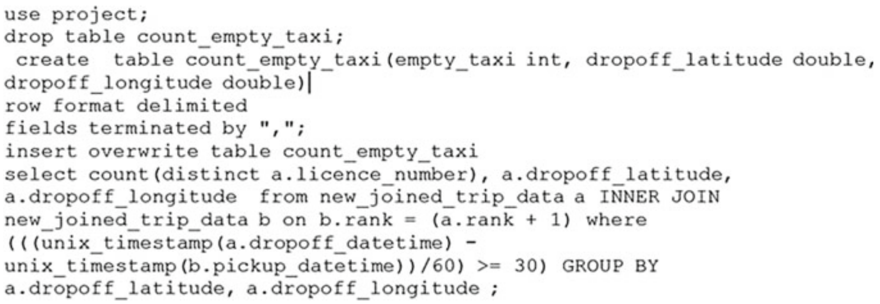
To analyze the empty taxis, a hive query is used to find the taxis whose time difference between the drop off time and next pick-up time is not more than 30 min and new hive table is created for this data count_empty_taxi. To analyze the revenue generated from an area, a hive query is written to find the amount collected by all the taxis within last 15 min. Self-join is performed and new hive tables is created count_profit. For a profitable area the profit should be more and the count of empty taxis should be less so a new hive table count_profitable_areas that is depicted in Fig. 1 is created for that data by performing inner join on above Fig. 1. The required result is transferred into MySQL database using sqoop which transfers the data from HDFS to RDBMS and vice versa. Finally, the MySQL is connected to web application which uses google maps API and top 10 profitable locations are displayed.

4 Results and Discussion

The pseudocode presented in Fig. 2 are used to identify the profit generated across different locations. Table 1 summarizes the output of the work by providing the projected profit (in $) along with the exact locations (in terms of Latitude and Longitude).

Table 1 Output of profits generated in the locations

Profit (in $)	Latitude	Longitude
86.5	40.56864547729492	−74.14933013916016
212.0	40.56901168823242	−74.34701538085938
56.5	40.574188232421875	−73.97637176513672
47.5	40.57433319091797	−73.99626159667969
43.5	40.57615661621094	−73.96410369873047
60.5	40.57645034790039	−73.95732879638672
59.5	40.57683563232422	−73.95315551757812
44.0	40.57704162597656	−73.96144104003906
52.0	40.5774040222168	−73.9795913696289
38.0	40.57807922363281	−73.9572525024414
54.5	40.57920837402344	−73.93714141845703
55.5	40.579593658447266	−74.00267791748047
50.5	40.58046340942383	−73.9678726196289
51.0	40.580528259277344	−73.96781921386719
38.0	40.58066940307671	−73.96776580810547



```
use project;
drop table count_empty_taxi;
 create  table count_empty_taxi(empty_taxi int, dropoff_latitude double,
dropoff_longitude double)
row format delimited
fields terminated by ",";
insert overwrite table count_empty_taxi
select count(distinct a.licence_number), a.dropoff_latitude,
a.dropoff_longitude  from new_joined_trip_data a INNER JOIN
new_joined_trip_data b on b.rank = (a.rank + 1) where
(((unix_timestamp(a.dropoff_datetime) -
unix_timestamp(b.pickup_datetime))/60) >= 30) GROUP BY
a.dropoff_latitude, a.dropoff_longitude ;
```

Fig. 3 Count of empty taxis in the location

The pseudocode presented in Fig. 3 are used to identify the number of empty taxis across different locations. Table 2 summarizes the output of the work by providing the count of empty taxis in number along with the exact locations (in terms of Latitude and Longitude) (Fig. 4).

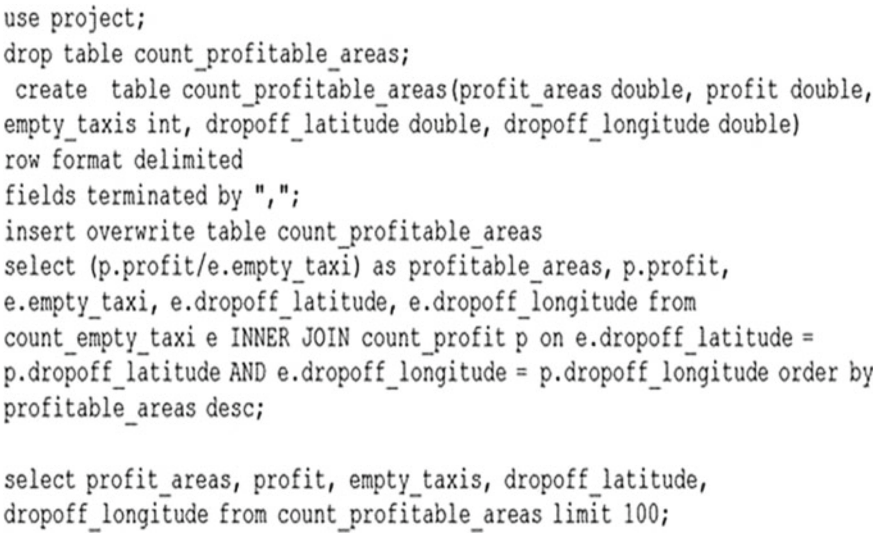
Figure 5 displays the top 10 profitable locations of New York City in a web application. Table 3 summarizes the profitable locations along with the number of empty taxis and revenue generated in that location.

Table 2 Output of number of empty taxis and locations

Number of empty taxis	Latitude	Longitude
1	40.56864547729492	−74.14933813916816
1	40.5743331989197	−73.99626159667969
1	40.57807922363281	−73.9572525024414
1	40.57920837482344	−73.93714141845703
1	40.58046340942383	−73.9678726716289
1	40.58920659555664	−73.93878173828125
1	40.59996795654297	−73.98482513427734
1	40.60086441040039	−73.99456024169922
2	40.60153579711914	−73.95244598388672
1	40.60184097290039	−74.07186889648438
1	40.60205841064453	−73.99613952636719
1	40.604549407958984	−73.97441864013672
1	40.6048583984375	−73.97329711914062
1	40.60639572143555	−73.91329956054688
1	40.60639857213356	−74.01895896415747
1	40.607200622558594	−74.00898742675721
1	40.60774612426758	−73.96766662597656
1	40.609012603759766	−73.97135925292969
1	40.613338470458984	−73.98967742919922
1	40.6148681640625	−74.03422546386719

Table 3 Latitudes and longitudes of the profitable locations

Profit areas	Profit	Empty taxis	Drop-off latitude	Drop-off longitude
230	230	1	41.0517196655273	−73.5350646972656
210	210	1	40.6338340759277	−73.597526550293
180	180	1	41.0193977355957	−73.6289978027344
123	123	1	40.6565284729004	−73.794242858867
111	111	1	40.6631813049316	−74.2363316650391
102	102	1	40.6951865119902	−74.1774978637695
102	102	1	40.7195472717285	−74.0321426391601
100	100	1	40.6878967285156	−74.18300628G 6211
100	100	1	40.9374542236328	−74.0181503295898
91	91	1	40.7361907958984	−74.3860244750977



```
use project;
drop table count_profitable_areas;
 create  table count_profitable_areas(profit_areas double, profit double,
empty_taxis int, dropoff_latitude double, dropoff_longitude double)
row format delimited
fields terminated by ",";
insert overwrite table count_profitable_areas
select (p.profit/e.empty_taxi) as profitable_areas, p.profit,
e.empty_taxi, e.dropoff_latitude, e.dropoff_longitude from
count_empty_taxi e INNER JOIN count_profit p on e.dropoff_latitude =
p.dropoff_latitude AND e.dropoff_longitude = p.dropoff_longitude order by
profitable_areas desc;

select profit_areas, profit, empty_taxis, dropoff_latitude,
dropoff_longitude from count_profitable_areas limit 100;
```

Fig. 4 Pseudo code for identifying profitable areas

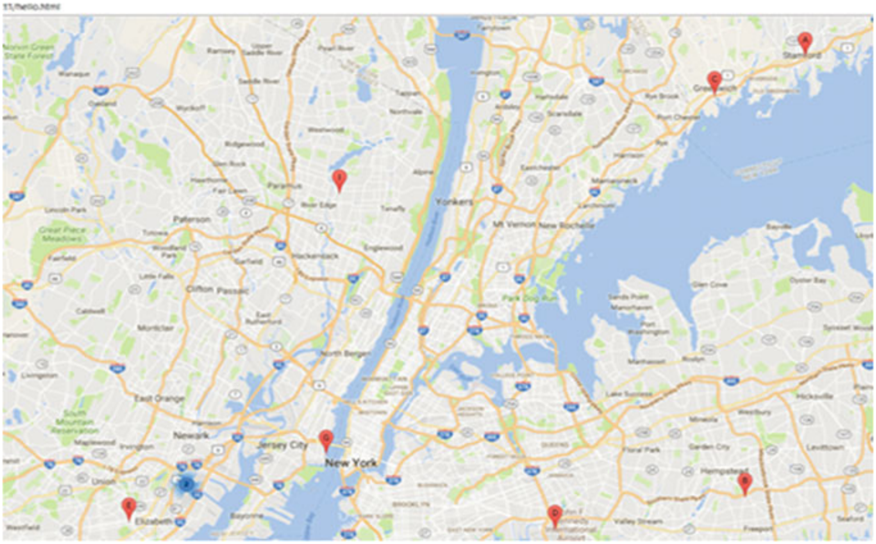


Fig. 5 Output displayed in the web application

5 Conclusion and Future Scope

In conclusion, the top 10 profitable locations in the New York City are displayed in the web application. Hadoop was used to analyze such a huge data in order to find the profitable locations. In the future, real-time data can be analyzed and the top locations which are nearby the drivers can be displayed. We can even extend this web application as a mobile application for both Android and IOS in future.

References

1. Shvachko, K., Kuang, H., Radia, S., Chansler, R.: The Hadoop distributed file system. In: Proceedings of 26th Symposium on Mass Storage Systems and Technologies (MSST), pp. 1–10. IEEE (2010)
2. Patel, U., Chandan, A.: NYC taxi trip and fare data analytics using BigData. Analyzing taxi data using bigdata, Oct 2015
3. Sun, H., Mcinstosh, S.: Big data mobile services for New York City taxi rides and drivers. In: IEEE International Conference on Mobile Services (2016)
4. Aslam, J., Lim, S., Pan, X.: City-scale traffic estimation from a roving sensor network. In: SenSys, Toronto, ON, Canada, 6–9 Nov 2012

Relevance Feedback Mechanism for Resolving Transcription Ambiguity in SMS Based Literature Information System

Varsha M. Pathak and Manish R. Joshi

Abstract This paper is about the development of an effective methodology that applies Relevance Feedback Mechanism to improve the relevance ranking of the documents for the noisy queries in an SMS based Information Systems in the domain of Literature of Indian languages. The system deals with the natural language queries which are freely formed by users in their own language with respect to literature related information access. The required flexibility in Romanized transliteration of the native scripts is considered in this problem. This paper presents the experiments and corresponding results of Marathi and Hindi language literature information system. We experimented on Marathi and Hindi literature which include songs, gazals, powadas, bharud and other types in a standard transliteration form like ITRANS. The ITRANSed literature documents are collected from Internet sites have semistructured format like latex documents. These documents are processed to build an appropriate Information Retrieval mechanism by customizing the conventional Vector Space Model. The paper presents the innovative methodology and proposed algorithms which resolves the ambiguity due to the noise in transcription method flexibly applied by user end. The proposed work applies Relevance Feedback Mechanism based on Rochchio's relevance classification. In this method user's preferences are considered to revise the relevance scores of top ranked documents delivered to them. Each document is identified as relevant or non-relevant according to user's judgments. This paper discusses these experiments. The results are obtained through a number of iterations by applying a sequence of relevance feedback gaining, processing and query refinement mechanism. Mean Reciprocal Rank is used to compare these results and evaluate the performance of our proposed model over a set of total 74 queries.

V. M. Pathak (✉) · M. R. Joshi
KCE Society's Institute of Management and Research Jalgaon (MS) India, North Maharashtra University, Jalgaon, India
e-mail: pathak.vmpathak.varsha@gmail.com

M. R. Joshi
e-mail: joshmanish@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_56

1 Introduction

Short Message Service (SMS) is a popular medium of information exchange on mobiles. It is the most economic way information access. It can be called as a reliable bridge for digital divide problem caused by web technology. From the integration of mobiles, Internet and other Information and Communication Technologies together, a ubiquitous world has almost emerged [1]. All parts of human lives would eventually be ubiquitously connected in future [2]. Mobiles and its underlined power of data store, retrieval and exchange using SMS service would play a key role in this changing world. Mobile industry and the telecommunication authorities are promoting these initiatives in the form ov Mobile Value Added Services (MVAS) [3].

1.1 Motivation

The possibility of using SMS for communication with a service running on a computer machine through a communication channel has emerged into an innovative research field termed as SMS based Information Retrieval (SMSbIR). A number of application areas and corresponding functionalities regarding SMS based information access are studied [4]. The respective literature includes a few government [5], banking [6], education [7] and health sectors [8]. These are the most popular areas in research. A few articles are also available on topics like SMS based tourism information access [5], SMS based home appliances controlling systems [9]. The mobile technologies like GSM, GPS, GPRS [10] are used in these applications.

In India the spoken languages and their respective scripting are two important features of “Personalization of Information” problem. With increasing digitization of libraries, literatures of various local languages from different states, are being restored in digitized form by applying innovative transliterations and encoding formats.

The Relevance Feedback Mechanism (RFM) is about recording the user’s reactions on the response of system for a query. The query and hence the results are improved by reevaluating the relevance of top ranked documents as per user’s reactions.

Our interest is to develop an appropriate retrieval mechanism with support of an effective RFM for relevant information access for the SMS based Literature Information System. We want to develop a system that receives an SMS query in transliterated form with respect to literature of Indian languages and sends back appropriate answerable document contents to the sender by using the same SMS channel. We have proposed a retrieval mechanism for Marathi and Hindi literature which represented by a collection of ITRANS documents.

This Introduction Section is followed by Sect. 2 which states the problem definition and its background concepts. Following this Sect. 3 specifies the document retrieval policy for noisy queries. The proposed algorithm of related work is discussed in this section. Section 4 explains the details of the proposed Relevance Feedback

Mechanism developed to improve the response of our system. It also presents Results and Evaluation of models using standard metrics. Section 5 is the Conclusion Section.

2 Problem Identification

As Indian languages vary in many respect from region to region, dealing with this variation in itself is a challenge. Scripting differences of Indian regional languages restrict the information exchange amongst the people living in diverse geographical area. To contribute in resolving this diversification we have chosen Literature Access in Indian languages by applying standard transliteration such as ITRANS. Lyrics of Hindi songs and related information like singer of the song, title, film of the songs are the most popular type of Hindi queries those are curiously asked by the users. As part of the development of an effective method of relevant information access for an SMS based Literature Information Systems, we dealt with Marathi and Hindi Literature Information Access using transliterated queries. We named our system as **Mobile Library** (MobiLib) that allows users to access literature related information by sending appropriate queries. To allow flexibility in transliteration we have worked on ambiguity in the transliterated queries due to variation in transcription encoding style. Our problem considers that the Romanized queries created by users may not follow a standard transliteration style. Similarly while typing a query in transliterated form errors are induced due to varied SMS style, misspelled terms.

This problem is thus specified under Noisy Channel assumption as discussed in [2].

2.1 Problem Definition

We identify this problem in novel manner as discussed below.

If Q is a Normal Query with some normal terms $t_j \in V$, where V is the vocabulary of specified information system. Due to transcription error on the channel, the query Q is distorted to a corrupted query Qc. So the problem is to map all the distorted words $tc_i \in Qc$ to corresponding normal terms $t_j \in Q$. Ultimately the challenge is to refine noisy query Qc in a relevant query RQ such that applying RQ will produce the same result as applying the Normal Query Q. Thus a user query Qc is considered as the combination of distorted terms, such that $Qc = tc_1 \cup tc_2 \in \dots tc_n$. This assumes that for some distorted term tc_i, there is a closely relevant normal term $t_j \in V$. Thus mapping tc_i to a most relevant normal term is the problem defined as term normalization. Based on this approach we want to test the hypothesis that, a maximum number of user intended terms are predicted in this refined query set RQ such that the probability of the relevant documents ranked at higher position is more than that of non-relevant documents. Further we apply Relevance Feedback Mechanism to improve the results such that the most relevant document shall occur at first position. Any further reference to the same query is thus accelerated in future.

2.2 About Dataset

The Marathi corpus is consisting of diverse categories of Marathi literature that mainly includes larger sized Spiritual literature like daasbodh, dnyaneshwari, geetarahasya etc. Similarly different small sized documents mainly cover bharud, kavita, songs, bhondalyache gaane, powadas. In Hindi corpus single category of literature is represented by cinema songs which is the most popular type of Literature. These documents are available in special encoded format which uses a standard transliteration popularly known as ITRANS. ITRANS is the abbreviation of Indian Language Transliterate. The ITRANS encoded Marathi Literature and Hindi Song collections are downloaded from the websites "http://www.sanskritdocuments.org" and "http://www.aczoom.com" to build the corpus of the system.

More than one thousand songs are processed to constitute the Hindi Vocabulary in the form of Inverse Document Index. Near about **fourteen thousand six hundred** terms form the complete Hindi song Vocabulary. Comparatively only one hundred and fifty Marathi documents which cover variable sizes of literature produce total **twenty two thousand seven hundred forty eight** terms. These terms are indexed to constitute Marathi vocabulary.

2.3 Related Work

The related literature shows that at major the research work in Aw et al. [11] which introduced the term SMS Normalization is the basic concept. This basic model applies Word Normalization to normalize the user query. According to this theory SMS Normalization is defined as "Removing of noise from SMS with a process that substitutes non-standard text (words) with standard text (words)". The non-standard words have noise like inclusion of digits, acronyms and spacing errors [11]. A statistical model is developed to deal with such problems [12]. A phonetic based algorithm to translate SMS message to original normal English language sentences is suggested by Pinto et al. [13]. Similar approach with two phases of SMS normalization is applied in [14]. These approaches have considered the SMS language as sub-language of the normal natural language.

Acharya et al. [15], have identified Multilingual phrasing and Misspellings as the major sources of errors in the SMS queries. Chen and his colleagues [2], have worked on a system that they call as SMSFind. The corresponding system uses Internet search engine(s) to extract relevant web pages for the given SMS query. They have highlighted in their work the application of Vector Space Model in searching algorithms.

SMS based Frequently Asked Question Systems (SMSbFAQS) is another vibrant topic under the domain of SMSbIR. This type of system is supported with a set of predefined queries. All these queries are constructed from the possible questions of specified domain or collected from some source such as a search engine, call centers or a web site. The nearest FAQ query is chosen for the SMS query sent by the user. This task requires IR strategy to normalize SMS query and apply similarity measures to map user query to one or more standard FAQs from the set. We have studied the state-of-the-art with respect to SMS based FAQ Systems and various issues by referring related work available in a few research papers. According to Kothari [16], a Frequently Asked Question (FAQ) system is designed by combining the similarity formulation of a searching mechanism to match the user query with the nearest FAQ sample query in the predefined query set (FAQs). In this related work SMS normalization technique is based on the Levenshtein's edit distance algorithm. This algorithm applies similarity measures based on consonant skeletons of the SMS tokens and FAQ Query term. The similarity score of a FAQ is calculated as the sum of similarity weights over all terms in the standard query Q for the corresponding user query S. An unsupervised approach is applied for automating this FAQ selection method. Mhaisale et al. [17] has modified this work by applying certain weights for pruning the Longest Common Subsequence ratio. Waghmare and Potey [18] have classified the SMSbFAQS as Monolingual, Cross-lingual and Multilingual where the linguistic computational complexity is in the increasing order.

The shared tasks of FIRE also address at related problems in SMS Query Translation, Query normalization in European and also in Indian languages like Marathi, Hindi, Bengali and many others [19]. Transliterated Search is addressed as the accepted solutions by the researchers working on Mixed Script and Cross Language Information Retrieval problems [20].

3 Noisy Query Similarity Mapping

In order to represent the corpus on our system Vector Space Model (VSM) is customized to represent the transliterated documents of Marathi and Hindi literature. Appropriate term weighting algorithm is applied to build the document term vector. The terms having weights greater than a threshold weight which is empirically considered as 0.5, are selected to build the Vocabulary V of the system. The term vector is the inverse document term index in which the terms occurring in the documents are organized with corresponding weights in the documents. The weights are calculated using **Term Frequency Inverse Document Frequency** algorithm (TF-IDF) as expressed by the Eq. 3. The term frequency is computed by applying the formula in Eq. 1.

$$tf(t,d) = \frac{freq(t,d)}{Maxfreq(t,d)} \tag{1}$$

$$idf(t) = \log(N/n) \tag{2}$$

$$tfidf(t,d) = tf(t,d) \times idf(t) \tag{3}$$

In these expressions, t is the term number, d is document number, N is the number of documents in the corpus and n is the number of documents in which the term t occurs.

For query collection, an interface was designed. The participants of corresponding experiment were instructed to send Devanagari queries in Roman Transliterated form. Freedom of transliteration style, SMS style was given to understand the possible variations. Around thirty different Devanagari handwritten queries were provided to them for reference. Near about 250 queries were collected which mostly reflect the variation in transliteration and SMS formation style.

Randomly selected thirty Marathi queries and forty four Hindi Queries from the query set were experimented to test the system's responses.

$$ProxWt_{t,v} = PrunWt - \frac{ed_{t,v}}{len} \tag{4}$$

where

- PrunWt is the pruning weight computed by applying certain rules as discussed in **Term Normalization Rules**.
- ed is the minimum edit distance calculated by Levenshtein's function minDist(t,v).
- length is considered as the length factor by applying term length rules.

3.1 Retrieval Policy

The important contribution of our model is that we proposed a customized Vector Space Model to build the document term vector by processing a collection of ITRANS documents. In this model instead of normalizing the user query to a nearest matching query occurring in the predefined set of Normal Queries, our model normalizes individual terms by selecting a set of Normal Terms from the system's vocabulary. These terms are at closer edit distance from the query term. For each Noisy term $tc_i \in Qc$ and a Normal Term $t_j \in V$ the minimum edit distance $ed_{ij} = tc_i \simeq t_j$ is computed by combining a well defined set of rules with the Levenshtein's minimum edit distance algorithm. Most of the related work of term normalization is based on Levenshtein's edit distance. The term normalization algorithms as in [16, 17] consider consonant skeleton comparison for computing the similarity between two words. In Marathi and Hindi language two totally different words can have same consonant skeleton. For example lataa and laataa, kavita and kuvata, lalitaa and lalita.

3.2 Term Normalization Rules

Considering the above discussions a set of rules are identified by understanding this kinds of language specific discriminations. These rules are varied to identify seven different rule based models. From these seven rule based models a best performing Rule based Model is chosen for the system. The corresponding rule set is defined in this subsection.

For defining this rule set it is considered that the SMS query in Qc is a sequence of n number of terms $tc_1, tc_2, \dots tc_n$. It is clear that the query is noisy as one or more terms are distorted due to noise in transcription and SMS encoding mechanisms.

Let V is the vocabulary constructed by processing the collection of ITRANSed literature documents as discussed above by applying our customized Vector Space Model. The terms $t_1, t_2, \dots t_m$ are the normal terms as they occur in V. ed = $tc_i \simeq t_j$ = minDist(tc_i, t_j) is the function that calculates minimum edit distance between two terms tc_i and t_j. l_1 and l_2 are the lengths of two strings normal term t_j and noisy term tc_i respectively. And let $W_1 = 0.60$, $W_2 = 0.40$, $W_3 = 0.2$, $W_4 = 0.75$ and $W_5 = 0.25$ are the weights used in the rules for Marathi literature. For Hindi corpus these weights are set as 0.85, 0.15, 0.5, 0.75 and 0.25 respectively. Empirically these weights are fine tuned in the term normalization experiments by varying them from 0 to 1.

Using above data following are the set of rules defined from our experiments. Rule 1 to Rule 5b are the length related rules and Rule 6a to Rule 9c are the weight related rules.

Rule 1: The first rule says that l1 $>$ 1 and l2 $>$ 1 is essential precondition for proxy weight calculation.
Rule 2: Compute length = $l_1 \geq l_2 \; ? \; l_1 : l_2$.
Rule 3: For a vocabulary term $t_j \in V$, if $l_1 \geq l_2$ then length = l_1 Vocabulary term t_j is qualified for further process of Proxy Weight calculation.
Rule 4: Compute length = $\frac{l_1+l_2}{2}$.
Rule 5a: Let threshold $\alpha = \frac{length}{2}$.
Threshold condition is $ed \leq \alpha$. To compute Proxy Weight for the term t_j, with respect to query term tc_i, the threshold value of edit distance ed(tc_i, t_j) is equal to half of the length, where length is the term length calculated applying length rules as above.
Rule 5b: Rule 5a is varied to $\alpha = \frac{length}{3}$
Following are the weight related rules.
Rule 6a: if first characters of tc_i and t_j match then flag1 = 1, wf1 = W_1.
Rule 6b: if first characters of tc_i and t_j do not match then wf1=W_2.
Rule 6c: if first characters of tc_i and t_j do not match then wf1=0.
Rule 7a: if second characters of tc_i and t_j match then wf2 = W_2.
Rule 7b: if second characters of tc_i and t_j do not match then wf2 = W_3.
Rule 7c: if second characters of tc_i and t_j do not match then wf2 = 0.
Rule 8a: if last consonant characters of the terms tc_i and t_j match then flag2 = 1, wf3 = W_4.

Rule 8b: if last consonant characters of the terms tc_i and t_j do not match then wf3 = W_5.
Rule 8c: if last consonant characters of the terms tc_i and t_j do not match then wf3 = 0.
Rule 9a: if threshold condition is true then apply following condition
if (flag1 == 1 and flag2 == 1) then Pruning Weight = 1.
Rule 9b: if Not (flag1 == 1 and flag2 == 1) then Pruning Weight = wf1 + wf2 + wf3 + wf4 + wf5.
Rule 9c: if threshold condition is true then unconditionally apply Pruning Weight = wf1 + wf2 + wf3 + wf4 + wf5.

3.3 Query Refining Algorithm

Consider that Qc is the noisy query submitted by the user with intention of retrieving information from the literature collection. If the query Qc contains n number of terms then possibly some or all of its terms are distorted. It is assumed that each term in the vocabulary V is normal term and is a single term query. Hence we modify the problem of FAQ system into and segment a query Q in a set of single term queries as $t_1, t_2, \ldots t_n$. We refine each single term of noisy query $t_i \in Q$ by substituting it with a relevant normal term set RQ_i. The set of candidate normal terms is computed by applying the rule based term normalization method. And finally all refined normal query term sets 'RQ_i's are combined to formulate a intentional query set. Where the intentional query $RQ = RQ_1 \cup RQ_2 \cup \cdots \cup RQ_n$. The algorithm Refine(Q, RQ) is thus developed to compute intentional relevant query set RQ for the user's noisy query Q. Following is its definition. The RQ set resulted by applying this algorithm for a sample query is presented in Table 1.

3.4 Precision Value Analysis

Different models are applied by varying the rule set. Out of these experimental models seven better performing models are evaluated by using precision measure. We term these variations as Models M1 to M7 respectively. Most of the queries are answerable by top ranked twenty documents. For the queries which are not answerable by any of the top 20 ranked documents are marked as NA. If a query is not answerable by any of top 20 documents the problem is resolved by applying other rule set with diluted constraints. Out of the seven models in most of the queries Model M7 and M4 are proved to be better performing. Model M7 is chosen for further processing for Marathi as well as Hindi corpus. The performance is measured by taking average precision values over all queries in sample data set. For example if for 6 out of 30 queries, their most relevant document is ranked at position 1 then the Precision

Table 1 Noisy term to normal term mapping

S. No.	Noisy term (tc)	Normal term (RQ_i)	Proxy weight (PW)	Doc#	Wt.	Doc#	Wt.	Doc#	Wt.
1	hamsfr	hamasafar	0.666666667	14	4.33073334	62	4.33073334	88	8.661466681
1	hamsfr	hamaar	0.666666667	215	6.906754779				
1	hamsfr	hamapar	0.571428571	285	6.906754779				
1	hamsfr	haar	0.5	16	3.401197382	114	6.802394763	140	3.401197382
1	hamsfr	hazaar	0.5	21	4.127134385	143	4.127134385	158	4.127134385
2	ltaa	lataa	0.55	28	24.04399911	101	8.01466637	113	4.007333185
2	ltaa	luTaa	0.55	141	8.25426877	145	8.25426877	150	4.127134385
2	ltaa	letaa	0.55	932	13.81350956				
2	ltaa	laa	0.5	27	7.783640596	65	7.783640596	160	3.891820298
2	ltaa	llaa	0.5	472	6.906754779				
3	aadmi	Aadmi	1	67	6.906754779				
3	aadmi	aadamii	0.714285714	115	4.418840608	166	4.418840608	238	4.418840608
3	aadmi	aadhii	0.666666667	38	5.111987788	499	5.111987788	687	5.111987788
3	aadmi	aa.Dii	0.666666667	238	13.81350956				
3	aadmi	Aandhi	0.666666667	767	6.906754779				
4	filmse	figure	0.5	4	6.906754779				
4	filmse	fikare	0.5	246	6.906754779				
4	filmse	firate	0.5	700	6.906754779				
4	filmse	fiqare	0.5	883	6.906754779				
4	filmse	farishte	0.125	334	6.212606096				

Algorithm 1 Refine(Q, RQ) : Qc is the input noisy query and RQ is the predicted output of user intended normal query term set.

```
Require: $n \geq 0 \vee x \neq 0$
Ensure: $RQ \subset NQ \subset V$
  1.Read Query Qc, assuming Qc is noisy, as $tc \in Qc$ may be noisy.
  2.Initialize
  $RQ \leftarrow \emptyset$
  3. Apply following steps.
  $NT \leftarrow \emptyset$
  4.$\forall tc \in Qc$ follow steps 5 thru 7
  5.$\forall t_i \in V$
  Compute Proxy Weight applying Equation 4 with weight rule set.
  $pw(t_i) \leftarrow tc \simeq t_i$
  $NT \leftarrow \cup t_i$
  6.Rank the terms in NT
  $\forall t_i \in NT$
  Rank the terms $t_i$ such that for any two terms $t_i, t_j$
  if $pw(t_i) \geq pw(t_j)$ then
     $Rank(t_i) \leq Rank(t_j)$
  end if7. Select top 5 ranked query terms nt($nt_1$, $nt_2 ... nt_5$) and add them in RQ
  $RQ \leftarrow RQ \cup nt$
  8. Output RQ for document relevance ranking step.
```

Value p@1 is equal to $\frac{6}{30}$. This way the average precision values are computed for first position as p@1. In similar manner precision values are measured at the rate of 1, 5, 10, 15 or 20 as p@1, p@5, p@10, p@15 and p@20 respectively. This discussion is presented in tabular form as Table 2.

4 Relevance Feedback Mechanism

Applying the Proxy Similarity Scoring function query intended by user could be predicted with higher probability than using the minimum edit distance method in raw form. The set RQ is assumed to be consisting of intended normal terms with maximum probability for a distorted query Qc. The cosine similarity function is used to formulate the proxy similarity score ProxSim. This algorithm is expressed in Eqs. 5 and 6 by considering the pruning weights of the normal query terms in RQ.

Here pw is the pruning weight calculated by our rule based selected model M7. If d_k is any document in the document set D, ProxSim is computed as the Cosine Similarity Score between user query Qc and the document d ∈ D. RQ is refined query computed by the proposed algorithm Refine(Qc, RQ). The corresponding Proxy Similarity computing function is defined in Eq. 6. As it is expressed in this equation Proxy Similarity between a distorted query Qc and a document d_k is computed by the Cosine Similarity between the RQ and the concerned document. This Cosine Similarity algorithm is stated in Eqs. 5 and 6.

Table 2 Evaluation of rule based transcription ambiguity resolving models

Model #	Description		P@1	P@5	P@10	P@15	P@20	NA
M1	Rule set	1,4,5b,9	0.273	0.727	0.864	0.955	0.955	0.05
	Number of queries	22						
	Number of Terms in RQ	84						
M2	Rule Set	1,2,5b,9	0.318	0.818	0.910	0.955	0.955	0.05
	Number of Queries	22						
	Number of Terms in RQ	88						
M3	Rule Set	1,2,5a,6a, 7a,9b	0.273	0.682	0.864	0.955	0.955	0.05
	Number of Queries	22						
	Number of Terms in RQ	82						
M4	Rule Set	1,3,5a,6a:c, 7a:c,8a:b,9a	0.367	0.733	0.9	0.933	0.933	0.07
	Number of Queries	30						
	Number of Terms in RQ	114						
M5	Rule Set	1,3,5a,6a:b, 7a:b,8a:b,9a:b	0.033	0.133	0.3	0.367	0.467	0.53
	Number of Queries	30						
	Number of Terms in RQ	125						
M6	Rule Set	1,3,5a,6a:c, 7a:c, 8a:b,9c	0.167	0.567	0.8	0.867	0.867	0.13
	Number of Queries	30						
	Number of Terms in RQ	125						
M7	Rule Set	1,3,5a,6a:c, 7a:c,8a:b, 9a	0.367	0.733	0.8	1	1	0.00
	Number of Queries	30						
	Number of Terms in RQ	124						

$$ProxSim(Qc, d_k) = Sim(RQ, d_k) \tag{5}$$

$$Sim(RQ, d_k) = \frac{\Sigma_{i=1}^{l} pw_{ki} dw_{kj}}{\sqrt{\Sigma_{k=1}^{n}(pw_{ki})^2}\sqrt{\Sigma_{k=1}^{n}(dw_{kj})^2}} \tag{6}$$

The system is tested on Marathi as well on Hindi document corpus. If a query is not answered by top 30 documents of the relevance order then the query is considered as not answerable by the system. Total thirty Marathi and forty four noisy Hindi queries were selected by applying appropriate sampling method. Some of the queries are answerable by multiple documents. For these queries the recall values are computed at top ten documents in the first iteration of query execution. The system shows low recall in some of the cases in these queries. On an average 0.49 recall value at the top ten documents is observed over all the 30 queries of Marathi literature. 0.37 is the average Recall value for Hindi cinema song corpus over all the 44 queries. For the queries those are answerable by single document recall is more than 0.91 for top 10 ranked documents. Considering all the cases system produces an average recall value 0.59 over all the 74 queries at top 10 ranked documents in first iteration.

4.1 Rocchio's Classification Model

Though the system shows good precision value, for effective use of a mobile based information access it is desirable to promote the most relevant document at first position. Also we need to improve the recall value of the multiple answerable queries. This ultimately requires to improve the relevance order by effectively promoting more relevant documents at higher ranks. For this purpose we have developed a Relevance Feedback Mechanism that modifies the similarity scores of the documents as per the relevance information supplied by users in implicit or explicit manners. Thus a systematic Relevance Feedback Mechanism (RFM) is proposed by applying the concept of Rocchio's classification [21]. The basic Rocchio's model is expressed by Eq. 7.

The RFM model of our system includes a systematic use of the mobile interface. The mobiles interface was first simulated using JAVA frames and then effectively it is incorporated in Android. The mobile client interface in Android is connected with the system service through Apache Tomcat server. This android user interface allows users to interact with the system in respect to the relevance of document list delivered by the system. At each iteration user looks at the snippets send by the system in a text box of the user interface for corresponding ranked document and feeds back whether the respective snippet is of answerable document or not.

The system processes this relevance feedback data received from user to modify the weights of terms in query RQ. As expressed in Eq. 7 the previous query is denoted by RQ_{Prv} and the revised query is denoted as RQ_{Rev}. This refined query term set RQ_{Rev} is then used to deliver modified top ranked document list on user's mobile interface.

$$RQ_{Rev} = RQ_{Prv} + \alpha \sum_{i=1}^{r} R_i + \beta \sum_{j=1}^{nr} NR_j \tag{7}$$

Through maximum five such iterations the most relevant documents get promoted at top rank in most of the cases.

The weights α and β are set as 0.8 and 0.2 respectively based on empirical observations.
If a term $t_i | R_i \in RQ \; \exists d \in D$, where d is given positive Relevance remark by user and R_i is its previous weight, then it is revised by adding 0.8× its value.
If a term $t_j | NR_j \in RQ \; \exists d \in D$, where d is given Non Relevance remark by user and NR_j is its previous weight, then it is revised by adding 0.8× its value.
Thus $\forall t \in RQ_{Prv}$ the weights are revised to refine RQ_{Prv} into RQ_{Rev}.
This refined query RQ_{Rev} is then used to re-rank the documents by applying the similarity algorithm as defined by Eq. 6.

4.2 Mean Reciprocal Rank Evaluation

For further evaluation of the retrieval method as discussed above, Mean Reciprocal Rank (MRR) is used. MRR is the standard measure used to evaluate retrieval systems with respect to its prediction about the first correct document for a sample query set [22]. It corresponds to Harmonic Mean of the ranks of first correct answer over all sampled queries. Equation 8 gives mathematical definition of MRR. For this all thirty Marathi queries and forty four Hindi queries are tested on our system by taking feedback of expert user. The user is provided with an appropriate GUI that displays the resultant document list at each iteration. The user needs to select correct document. The interface displays 4 lines snippets of visited documents. If a correct document is ranked at first position the system receives corresponding feedback from the user. At first run of each query the rank of first correct document for the respective query is taken into account. For Marathi Literature and corresponding query set 0.578 MRR value is calculated, where as Hindi Song Retrieval over all sample queries produces 0.442 MRR value.

$$MRR = \frac{\sum_{i=1}^{|Q|} \frac{1}{rank_i}}{|Q|} \tag{8}$$

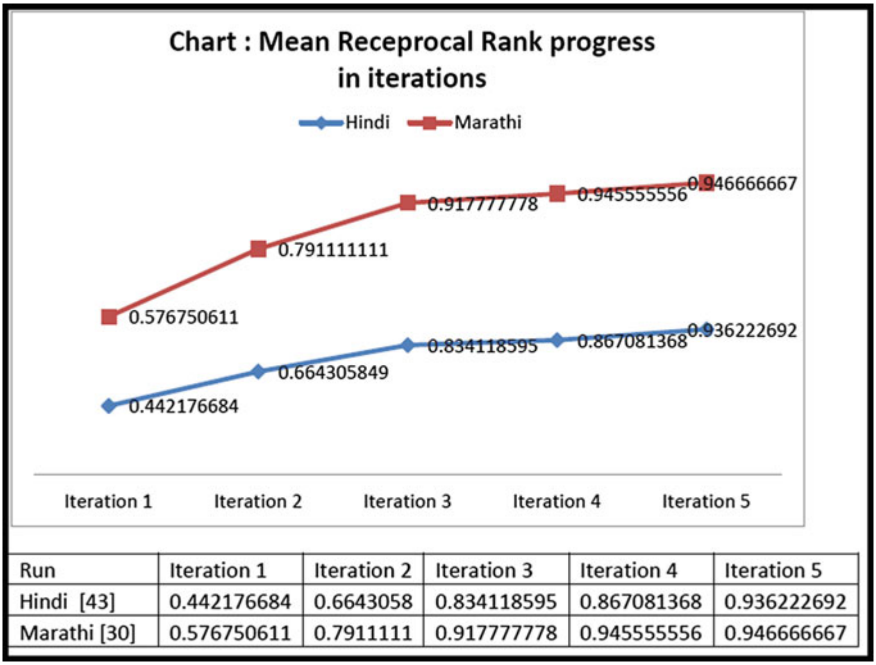
The above algorithm is applied on all thirty queries in Marathi and forty four Hindi queries using corresponding corpus. In case of Marathi queries, the system has given good results almost in all queries. Out of the seven selected rule based models, the model M4 and M7 are found producing appropriate relevance order. In most of the queries at first iteration using these Models all thirty queries are answerable by one of the top ranked 20 documents. The system's response is improved with application of RFM to achieve the aim of having most relevant document at first position. To judge this improvement MRR value is computed after every iteration. The MRR is improved from 0.577 at first iteration to 0.947 at the end of fifth iteration.

With the same measure of MRR, the system is evaluated over all of 44 Hindi queries by applying RFM as discussed above. Only one query is not answerable by any of the forty top ranked documents delivered by the system through planned five iterations. It is further confirmed that the specific document is not existing in the corpus. Thus system's response that, "**the query is not answerable**" at the end of fifth iteration is found correct.

For further investigation with respect to the system's performance on Hindi Song Retrieval, MRR is computed on remaining 43 Hindi sampled queries through all five iterations and found the results are improved. The MRR factor has increased from 0.442 of first iteration to 0.936 in final iteration.

This MRR evaluation is graphically presented in Fig. 1. This chart shows the comparison between Marathi and Hindi System.

Other important observation is that some of the queries are answerable by multiple documents. We computed the recall values for these queries. The recall value assures that maximum relevant answerable documents occur at top positions. The average recall of our system for these queries is almost 91%. All equally relevant documents occur at top positions in a sequence. The second ranking documents also occur in continuous sequence following the first ranked documents.



Run	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
Hindi [43]	0.442176684	0.6643058	0.834118595	0.867081368	0.936222692
Marathi [30]	0.576750611	0.7911111	0.917777778	0.945555556	0.946666667

Fig. 1 MRR evaluation of Marathi and Hindi literature retrieval

5 Conclusion

This paper is about the development and application of an effective Relevance Feedback Mechanism which resolves the ambiguity generated due to the transcription variations. The proposed algorithm normalizes a noisy query by normalizing the noisy terms to nearer normal terms. As a noisy term can match with more than single Normal Terms, a selection of top relevant terms is done by applying relevance ranking on normal terms. Top 5 such relevant terms for each noisy term are used refine the noisy query as whole. The relevance of normal term with user term measured by Proxy Weights calculation. Proxy Weight measures the closeness of the normal terms with the user query terms by applying a set of rules and the Levenshtein's minimum edit distance function.

The hypothesis that a relevance feedback mechanism can be effectively applied for improving the relevance ranking of documents is then tested. A Rocchio's classification model is implemented for refinement of the term weights on the basis of whether the term occurs in a relevant document or in a non-relevant document. The relevance judgments are taken by applying appropriate relevance feedback gaining method. Appropriate client server paradigm is developed to test performance of the developed system.

The Mean Reciprocal Rank shows an improvement in ranking from 0.577 to 0.947 for Marathi corpus. Similarly for Hindi corpus the improvement in ranking is observed from 0.442 to 0.936. Thus we conclude that the relevant answerable document can be promoted to the top rank position in relevance order by using appropriate relevance feedback processing and query refining methods.

References

1. Sundar, D.K., Garg, S.: M-governance: a framework for Indian urban local bodies. In: The Proceedings of Euro mGov, pp. 10–12 (2005)
2. Chen, J., Subramanian, L., Brewer, E.: SMS-based web search for low-end mobile devices. In: Proceedings of the Sixteenth Annual International Conference on Mobile Computing and Networking, pp. 125–136. ACM (2010)
3. Delloitte ASSOCHAM January 2011. Mobile value added services (mvas)—a vehicle to usher in inclusive growth and bridge the digital divide. Technical report
4. Joshi, M.R., Pathak, V.M.: A survey of SMS based information systems. arXiv preprint arXiv:1505.06537 (2015)
5. Chang-Jie, M., Jin-Yun, F.: Location-based mobile tour guide services towards digital dunhuang. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 37, issue B4, pp. 949–953 (2008)
6. Mousumi, F., Jamil, S.: Push pull services offering SMS based m-banking system in context of Bangladesh
7. Ismail, M.N., Mohd Nazri Ismail: Development of WAP based students information system in campus environment. Int. J. Comput. Theory Eng. **1**(3), 260–271 (2009)
8. Adesina, A.O., Nyongesa, H.O.: A mobile-health information access system (2013)
9. Raghavendran, G.: SMS based wireless home appliance control system. In: Proceedings of International Conference on Life Science and Technology (ICLST 2011) (2011)

10. Ramamurthy, B., Bhargavi, S., ShashiKumar, R.: Development of a low-cost GSM SMS-based humidity remote monitoring and control system for industrial applications. Int. J. Adv. Comput. Sci. Appl. **1**(4) (2010)
11. Aw, A., Zhang, M., Xiao, J., Su, J.: A phrase-based statistical model for SMS text normalization. In: Proceedings of the COLING/ACL on Main Conference Poster Sessions, pp. 33–40. Association for Computational Linguistics (2006)
12. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Vol. 1, pp. 48–54. Association for Computational Linguistics (2003)
13. Pinto, D., Vilarino, D., Alem, Y.: The soundex phonetic algorithm revisited for SMS-based information retrieval. Department of computer science, Mexico
14. Byun, J., Lee, S.-W., Song, Y.-I., Rim, H.-C.: Two phase model for SMS text messages refinement. In: AAAI Workshop on Enhanced Messaging (2008)
15. Acharyya, S., Negi, S., Venkata Subramaniam, L., Roy, S.: Unsupervised learning of multilingual short message service (SMS) dialect from noisy examples. In: Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data, pp. 67–74. ACM (2008)
16. Kothari, G., Negi, S., Faruquie, T.A., Chakaravarthy, V.T., Venkata Subramaniam, L.: SMS based interface for FAQ retrieval. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2, vol. 2, pp. 852–860. Association for Computational Linguistics (2009)
17. Mhaisale, S., Patil, S., Mahamuni, K., Dhillon, K., Parashar, K.: FAQ Retrieval Using Noisy Queries: English Monolingual Sub-task. DTU, Delhi, India, FIRE (2013)
18. Waghmare, J., Potey, M.A.: Domain specificity for focused retrieval in SMS based FAQ retrieval. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **5**(6), 1315–1321 (2015)
19. Banerjee, S., Naskar, S.K., Rosso, P., Bandyopadhyay, S.: The first cross-script code-mixed question answering corpus
20. Verulkar, P., Balabantray, R.C., Chakrapani, R.A.: Transliterated search of Hindi lyrics. Int. J. Comput. Appl. **121**(1) (2015)
21. Salton, G.: Automatic information organization and retrieval (1968)
22. Kowalski, Gerald: Information retrieval systems: theory and implementation. Comput. Math. Appl. **5**(35), 133 (1998)

Integration of Vector Autoregression and Artificial Neural Networks: A Robust Model for Prediction of Nonstationary Data

Akhter Mohiuddin Rather

Abstract This paper proposes a novel and robust prediction-based mathematical model. The proposed model is actually a hybrid of two different models: vector autoregression model and artificial neural network. The goal is to predict the patterns of any nonstationary data which is a big challenge; for this purpose, stock data has been chosen as the domain of this work. Since stock data is highly volatile and nonstationary, most of the linear models fail to capture its patterns. The proposed integrated model is a combination of a linear model and nonlinear model. Optimal weights are required to combine two different prediction-based models which have been generated using an optimization model and solved using genetic algorithms. The proposed integrated model is tested on nonstationary stock data and it is observed that the model is able to capture its patterns very well, resulting into excellent predictions.

1 Introduction

Prediction of nonstationary data such as of stock data has been a challenge for researchers, academicians as well as for industrialists. When it comes for linear data, traditional statistical models have been able to do well by predicting the future movements of such linear and stationary data. However, while predicting nonstationary data such as stocks, these linear models fail to meet up expectations. Therefore, researchers have started to find alternatives to overcome such problems. Artificial neural networks (ANNS) are very popular in the field of prediction-based system. ANNs contain some nonlinear functions known as activation functions which are able to detect the nonlinear movements of data very well. Prediction-based model are broadly classified into two categories, statistical models and Artificial Intelligence (AI) based models. Statistical models are also known as traditional models or linear models. Some popular statistical models include random walk model [1], linear

A. M. Rather (✉)
Vignana Jyothi Institute of Management, Bachupally 500090, Hyderabad, India
e-mail: dr.AkhterMR@vjim.edu.in
URL: http://www.vjim.edu.in

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_57

trend model [2], exponential smoothing model [3], the famous autoregressive integrated moving average models (ARIMA) [4] are known as one of the most incredible contributions in the literature of linear forecasting. ANNs are most commonly used AI-based models used for prediction of nonstationary data proposed decades back [5]. Support vector machines (SVM) are also well known AI-based models used in similar areas [6]. Many scholars have used ANNs along with linear models [7–11].

Time series models such as famous ARIMA models were first implemented on ANN by White [12], thereby such ANNs can also be called as autoregressive neural networks. ARIMA models were also implemented on four different ANNs and tested on stock data [10]. This particular area has gained momentum over a period of time [13, 14]. A unique work using autoregressive ANN was proposed [15] wherein the authors introduced prediction based optimal portfolio of stocks. Time series models have also been implemented on Radial basis function (RBF) and some scholars have claimed of getting better results of predictive analytics when using RBF [16–18].

Nonlinear models such as ANNs are expected to yield better results as compared to linear models, which doesn't turn out to be true always. ANNs do not always meet up to expectations, therefore some other methods or alternatives need to be explored. Integrating various forecasting models into a single powerful model is one such alternative. The resultant integrated forecasting model can also be called as hybrid forecasting model. The area of hybrid models or combined forecasting has received great attention from researchers, scholars, and industrialists from past one and half decades. The goal of forming any integrated model is very clear, i.e., it should outperform its subset of all individual models in terms of better results. Hybrid models can be formed by integrating various linear models, nonlinear models or linear and nonlinear models. The main drawback of ANNs is that they converge only once they reach up to minimum preset error threshold, which means time complexity is high. Therefore most of the researchers tend to include only one nonlinear model in hybrid system [19–21]. Recurrent neural network and ARIMA model was integrated and a robust model was formed, the model has been successfully experimented on stock data [22]. Another such integrated model was formed by combining random walk model and Elman ANN [23].

Rest of the paper is arranged as follows: Sect. 2 describes the main methodology used in this work, In Sect. 3 the proposed model is described in detail. Experimentation is shown in Sect. 4 and finally conclusions are presented in Sect. 5.

2 Methodology Used

This section provides an overview on the methodology used in this work.

2.1 Vector Autoregression

Vector autoregression (VAR) is a statistical measure where all the equations are treated endogenous [24]. VAR is an n-equation n-variable statistical model in which each variable is treated by its own lagged past observations. VAR model is an extension to univariate autoregressive model of time series. Hence, it forms regression on itself similar to autoregressive model. Suppose the length of a time series is T such that $(t = 1, \ldots, T)$ having T past observations. For the same time series, VAR model has set of k endogenous variables over the same period of time $(t = 1, \ldots, T)$ as a linear function of their past observations. The k endogenous variables are represented in a $k \times 1$ vector y_t, which has ith element as y_{it} the observation at time t of ith variable. A VAR model of order p denoted by VAR(p) can be expressed as shown in Eq. 1.

$$\mathrm{VAR}(p) = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \tag{1}$$

where α is $k \times 1$ vector of constants, ϕ_i is $k \times k$ matrix and ε_t is $k \times 1$ vector of errors. A simple VAR(1) be written as shown in Eq. 2.

$$\mathrm{VAR}(p) = \alpha + \phi_1 y_{t-1} + \varepsilon_t \tag{2}$$

VAR(1) with two with $k = 2$ variables can be written as shown in Eq. 3.

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} + \begin{bmatrix} \phi_{1,1} & \phi_{1,2} \\ \phi_{2,1} & \phi_{2,2} \end{bmatrix} + \begin{bmatrix} y_{1,t-1} \\ y_{2,t-2} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix} \tag{3}$$

Equation 3 has only one ϕ matrix because maximum lag in this case is $p = 1$. Constant α and ϕ can be determined using regression.

2.2 Artificial Neural Networks

ANNs are information processing units which have inspired from human brain [5]. In the field of prediction-based models and classification, ANNs have replaced existing linear models and they still continue to catch the interest of researchers. This is because of their excellent performance and robustness that they are still being explored and improved. Deep ANNs and Extreme learning machines are new and improved type of ANNs which have appeared in the literature [25, 26]. An ANN has an input layer, data is fed into ANN from input layer and output is collected from output layer. Nonlinear processing happens in a layer which is in between input and output layer, this layer is called hidden layer. Each layer consists of several neurons and each neuron processes data either in linear or nonlinear form. The power of ANN is in hidden layer, wherein data passes through nonlinear functions also known as activation functions.

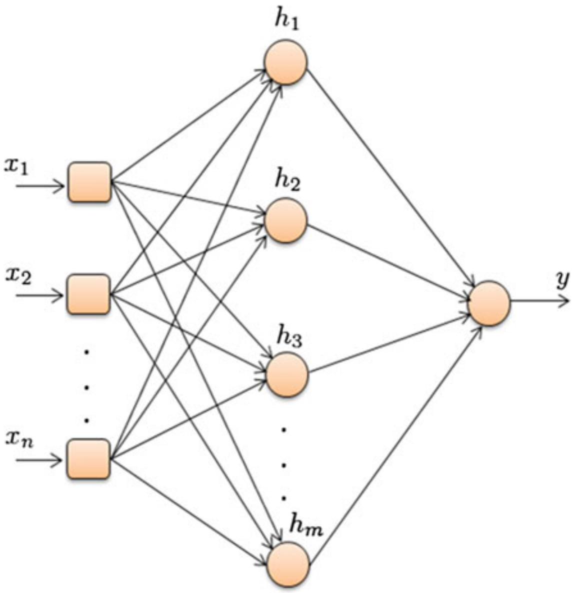


Fig. 1 Artificial neural network with one hidden layer

A typical ANN is shown in Fig. 1 shows a typical ANN with n input neurons $(x_1, x_2, \ldots, x_n)$, m hidden neurons $(h_1, h_2, \ldots, h_m)$. Such ANN is also known as multilayer perceptron (MLP) and is most widely used ANN. The connection links between different layers are associated with random weights and weighted sum is calculated at the forward layer. For instance, if the input neurons $x_1, x_2, \ldots, x_n$ are associated with weights w_1, w_2, ..., w_n, the weighted sum is calculated as shown in Eq. 4, which shows weighted sum being collected by jth hidden neuron.

$$h_j = a_0 + \sum_{i=1}^{n} x_i w_i \tag{4}$$

where a_0 is known as bias; final output y is collected after it passes through a nonlinear function $\sigma(y)$. A network learns based on the output shown to output layer, such learning is known as supervised learning. However, when no output is shown to output layer, in that case, the network learns on its own by unsupervised learning. The network has to go through several sets of iterations, such iterations are known as epochs. The network learns using backpropagation algorithm by adjusting weights in each epoch depending upon data, its characteristics such as how much noise it has, an ANN may require hundreds or millions of epochs to converge.

3 Proposed Integrated Model

Proposed Integrated Model (PIM) is obtained by integrating two different models: ANN which is a nonlinear stochastic model and a VAR model which is a linear model. The integration of these two models require optimal weights which have been generated using an optimization model, the model has been solved using Genetic Algorithms (GA).

Consider a nonstationary time series data $r_t(t = 1, \ldots, T)$. Its predictions are calculated separately from ANN and from VAR model. Let the weight vector for both models be, $W = [w_1, \ldots, w_2]$. Final prediction of PIM is obtained as shown in Eq. 5.

$$\hat{r}_t = \sum_{i=1}^{2} w_i \hat{r}_{it} \qquad (t = 1, \ldots, T) \tag{5}$$

The condition is that the weights should sum equal to 1 as shown in Eq. 6.

$$\sum_{i=1}^{2} w_i = 1 \tag{6}$$

Equation 7 shows the calculation of prediction error of PIM.

$$\varepsilon_t = r_t - \hat{r}_t \tag{7}$$

PIM can be thus obtained as shown in Eq. 8.

$$\hat{r}_{\text{PIM}(t)} = w_1 \hat{r}_{\text{ANN}(t)} + w_2 \hat{r}_{\text{VAR}(t)} \qquad (t = 1, \ldots, T) \tag{8}$$

where $\hat{r}_{\text{PIM}(t)}$ is prediction obtained from PIM, $\hat{r}_{\text{ANN}(t)}$ and $\hat{r}_{\text{VAR}(t)}$ are predictions obtained from ANN and VAR respectively.

The optimal weights of PIM have been obtained using the optimization model shown in Eqs. 9–11. The objective function of the model is to minimize the mean squared error (MSE) of predictions obtained from PIM. The two constraints specify that weights should range between 0 and 1 and also the weights must sum equal to1.

$$Min \quad \frac{\sum_{t=1}^{T} (r_t - \hat{r}_{\text{PIM}(t)})^2}{T} \tag{9}$$

$$s.t. \quad \sum_{i=1}^{2} w_i = 1 \tag{10}$$

$$0 \le w_i \le 1 \quad (i = 1, \ldots, 2) \tag{11}$$

4 Experimentation

The robustness of PIM is checked by experimentation on real data, for this purpose stock data has been chosen. Since stock market is volatile and the data is nonstationary, a large set of experiments were carried out so that robustness of PIM is checked. This section shows results of PIM on three noisiest stocks (out of several stocks), namely Andhra Bank, Idea Cellular and Jaypee Infrastructure. The daily returns of these stocks have been considered from September 23, 2014 to August 05, 2015 (164 returns), the data has been obtained from Bombay Stock Exchange (BSE).

4.1 Performance of ANN

Each stock of 164 observations has been divided in the ratio of 80:50; 80% of data is kept for training ANN and rest 20% for testing. Each stock is thereafter arranged as sliding windows; 33 sliding windows have been obtained for each stock. Each window gives one future prediction, therefore 33 windows give 33 predictions, and initial window gives prediction for future period (r_{t+1}). In each window, input data has been arranged in the form of autoregressive model of order 4, AR(4).

MLP[4:12:1] (4 neurons in input layer, 12 neurons in hidden layer, 1 neuron in output layer) has been chosen for the experimentation. Input layer has therefore four inputs as r_{t-1}, r_{t-2}, r_{t-3}, r_{t-4} fed into four corresponding neurons and target output r_t is shown to output neuron; the network is set for training in a supervised manner. Learning rate was set to 0.1 and momentum as 0.03. Before training, error threshold was set to 0.0002, the network converges only once the error reaches below the preset threshold. On average it took 150,000 epochs for MLP (for each stock) to reach up to the preset error threshold.

4.2 Performance of VAR

The performance of VAR has been minimal, due to its linear nature it was unable to capture the patterns of the data. This has resulted into high prediction error. However, it is expected that PIM outperforms ANN and VAR, which means both ANN and VAR contribute individually for better performance of PIM.

4.3 Performance of PIM

As desired, PIM outperforms both of its subset individual models, i.e., ANN and VAR. PIM is able to capture the nonstationary behavior of data very well. Figure 2

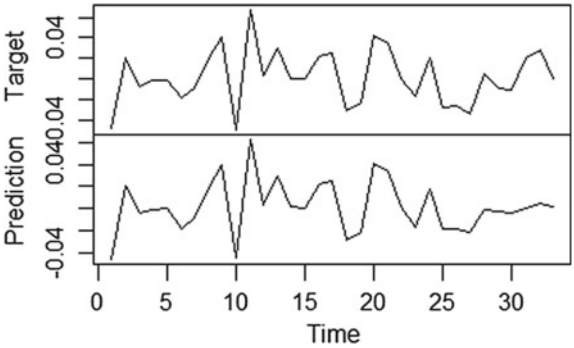


Fig. 2 Output of PIM

Table 1 Error metrics

	Andhra B.	Idea C.	Jaypee I.
VAR			
MSE	0.0007	0.0004	0.0007
MAE	0.0176	0.0170	0.0176
ANN			
MSE	0.0005	0.0002	0.0006
MAE	0.0270	0.0050	0.0152
PIM			
MSE	0.0002	0.0001	0.0002
MAE	0.0122	0.0040	0.0091

shows time series graph of target returns and predictions obtained from PIM for stock "Idea Cellular" As seen PIM is able to predict the fluctuations of nonstationary stock data very well. This results into least prediction error and target data seem to be closer to predicted data.

Table 1 shows error metrics of all three models (VAR, ANN and PIM) in consolidated form. The comparison of three models is therefore done by means of Mean Square Error (MSE) and Mean Absolute Error (MAE). It is observed that VAR has performed poorly leading to high prediction error. ANN results into less prediction error as expected. As desired, PIM outperforms ANN, in terms of least prediction error.

5 Conclusions

A new novel method of predicting nonstationary data was proposed here. For the sake of clarity stock data was considered as domain of the problem, the reason is that stock market is highly volatile, resulting into nonstationary behavior of data. However, the model is not limited to stock data only, it can be used and applied in other areas too. The limitations of the proposed model is that it is data dependent. The model can be further refined by applying some different regression schemes and including few more prediction based models (linear or nonlinear) in the integrated system. The future directions include to explore deep neural networks, extreme learning machines and feature selection for similar problems. This is certainly an important avenue for future research.

Acknowledgements Research support from University of Hyderabad, Institute for Development & Research in Banking Technology, Hyderabad and Vignana Jyothi Institute of Management, Hyderabad is highly acknowledged.

References

1. Malkiel, B.G.: A Random Walk Down Wall Street - The Time - Tested Strategy for Successful Investing. W. W. Norton & Co., New York (2008)
2. Altaya, N., Rudisillb, F.: Adapting Wright's modification of Holt's method of forecasting intermittent demand. Int. J. Prod. Econ. **111**, 389–408 (1981)
3. Brown, R.G.: Smoothing, Forecasting and Prediction. Courier Dover Publications, Mineola, NY (2004)
4. Box, G.E.P., Jenkins, G.M.: Time Series Analysis, Forecasting and Control. Holden-Day, San Francisco (1970)
5. McCulloch, W.S., Pitts, W.: A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. **5**, 115–133 (1943)
6. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**, 195–197 (1995)
7. Chen, C.H.: Neural networks for financial market prediction. In : IEEE World Congress on Computational Intelligence, pp. 1199–1202. IEEE Press, Orlando (1994)
8. Konno, H., Kobayashi, K.: An integrated stock-bond portfolio optimization model. J. Econ. Dyn. Control **21**, 1427–1444 (1997)
9. Konno, H., Yamazaki, H.: Mean-absolute deviation portfolio optimization model and its applications to Tokyo stock market. Manag. Sci. **37**, 519–531 (1991)
10. Rather, A.M.: A prediction based approach for stock returns using autoregressive neural networks. In : IEEE World Congress on Information and Communication Technologies, pp. 1271–1275. IEEE Press, India (2011)
11. Rather, A.M.: Computational intelligence based hybrid approach for forecasting currency exchange rate. In : IEEE 2nd International Conference on Recent Trends in Information Systems, pp. 22–26. IEEE Press, India (2011)
12. White, H.: Economic prediction using neural networks: the case of IBM daily stock returns. In : IEEE International Conference on Neural Networks, pp. 451–458. IEEE Press, New York (1988)
13. Chen, A.S., Leung, M.T., Daouk, H.: Application of neural networks to an emerging financial market: forecasting and trading the Taiwan stock index. Comput. Oper. Res. **30**, 901–923 (2003)
14. Jain, A., Kumar, A.M.: Hybrid neural network models for hydrologic time series forecasting. Appl. Soft Comput. **7**, 585–592 (2007)
15. Freitas, F.D., De Souza, A.F., de Almeida, A.R.: Prediction-based portfolio optimization using neural networks. Neurocomputing **72**, 2155–2170 (2009)
16. Hafezi, R., Shahrabib, J., Hadavandi, E.: A bat-neural network multi-agent system (BNNMAS) for stock price prediction: case study of dax stock price. Appl. Soft Comput. **29**, 196–210 (2015)

17. Kim, K.J., Ahn, H.: Simultaneous optimization of artificial neural networks for financial forecasting. Appl. Intell. **36**, 887–898 (2012)
18. Shen, W., Guo, X., Wu, C., Wu, D.: Forecasting stock indices using radial function neural networks optimized by artificial fish swarm algorithm. Knowl. Based Syst. **24**, 378–385 (2011)
19. Zhang, G.P.: Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing **50**, 159–175 (2003)
20. Kwon, Y.K., Moon, B.R.: A hybrid neurogenetic approach for stock forecasting. IEEE Trans. Neural Networks **18**, 851–864 (2007)
21. Wang, J.J., Wang, J.Z., Zhang, Z.G., Guo, S.P.: Stock index forecasting based on a hybrid model. Omega **40**, 758–766 (2012)
22. Rather, A.M., Agarwal, A., Sastry, V.N.: Recurrent neural network and a hybrid model for prediction of stock returns. Expert Syst. Appl. **42**, 3234–3241 (2015)
23. Adhikari, R., Agrawal, R.K.: A combination of artificial neural network and random walk models for financial time series forecasting. Neural Comput. Appl. **24**, 1441–1449 (2014)
24. Sims, C.A.: Macroeconomics and reality. Econometrica: J. Econ. Soc. **48**, 1–48 (1980)
25. Liew, S.S., Khalil-Hani, M., Bakhteri, R.: Bounded activation functions for enhanced training stability of deep neural networks on visual pattern recognition problems. Neurocomputing **216**, 718–734 (2016)
26. Soares, S.G., Arajo, R.: An adaptive ensemble of on-line extreme learning machines with variable forgetting factor for dynamic system prediction. Neurocomputing **171**, 693–707 (2016)

Clustering with Polar Coordinates System: Exploring Possibilities

Yogita S. Patil and Manish R. Joshi

Abstract Clustering is unsupervised learning technique to group similar objects. The quality of clustering is assessed by several internal as well as external measures such as Dunn index, Davies–Bouldin index (DB), Calinski-Harabasz index (CH), Silhouette index, R-Squared, Rand, Jaccard, Purity and Entropy, F-measures, and many more. Researchers are exploring different approaches to improve quality of clustering by experimenting with different partitioning strategies (similarity/distance formula), by changing representation of data points or by applying different algorithms. In our earlier research paper (Joshi and Patil in 2016 IEEE Conference on Current Trends in Advanced Commuting (ICCTAC), pp 1–7, 2016 [1]), we put forth our observations of changing coordinate system of objects from Euclidean to Polar on clustering. In continuation, we further experimented to explore the possibilities of clustering with different distance techniques for partitioning objects represented in Polar coordinate system. We experimented with a standard as well as real data set. The quality of clustering is evaluated using Silhouette internal evaluation measure.

1 Introduction

Clustering is an unsupervised learning technique to group homogeneous objects into one cluster so that all objects in one cluster similar as possible and dissimilar to objects in different clusters as possible. The goal of clustering is to explore the large amount of data and mine interesting patterns inside it. In the last 10 years, many researchers have proposed large number of clustering algorithms. As data is having different characteristics (like size, high dimensionality, noise, and outlier, sparseness, etc.) clustering methods are categories as, partitioning methods, hierarchical

Y. S. Patil (✉)
North Maharashtra University, Jalgaon, Maharashtra, India
e-mail: patilmyogita@gmail.com

M. R. Joshi
School of Computer Sciences, North Maharashtra University, Jalgaon, Maharashtra, India
e-mail: joshmanish@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_58

methods, density-based methods, grid-based methods, and model-based methods, probability-based, correlation based, and mixed attribute based [2]. All clustering algorithms use different distance measures to determine the similarity between various objects in data set. A number of different distance measures have been proposed to measures the distance between observations like Euclidean distance, Manhattan distance, Pearson, hamming distance, Cosine similarity, Jaccard index, etc. The data used for cluster analysis can be interval, ordinal, categorical, or mixed. An appropriate distance measure must be chosen based on the properties of data set in order to obtain correct clusters. The type of distance measures used can affect the cluster analysis likewise a better representation of data points could yield better clustering result [3]. In this paper, we experiment with two different distance formula in order to test the outcome of alteration in distance formula on clustering with Polar coordinates.

In our earlier research paper [1], we conducted a study to identify the result of the change in coordinate system on clustering. We proposed clustering algorithm to group objects represented using Polar coordinate system [1]. To get a better clustering solution for the data points which are not separable by Euclidean distance, one can choose Polar representation of same points [4]. For our research, we mapped Cartesian coordinates to the Polar coordinate system. The concept of Polar coordinates and mapping of Cartesian coordinates to Polar coordinates is well described in [1]. We have used the commonly used Euclidean distance formula to measure the likeness between two objects represented in Polar coordinate system. Based on experimental result, we observed that for certain non-globular shape cluster change in coordinate system, i.e., Cartesian to Polar gives better result. We compare the performance of conventional k-means for Cartesian coordinates as well for Polar coordinates and our proposed algorithm [1] using several cluster validity measures. Consequently, we conclude that a proper illustration of data points may yield a proper clustering result.

Continuing the same study in this paper, we aimed at exploring the possibilities of clustering with Polar coordinates using two different distance formulas, i.e., Euclidean and Polar. In order to achieve the goal, we compare the outcome of distances formulas in Polar coordinate system on clustering using Silhouette internal evaluation measure. The commonly used Euclidean distance formula (Distance Formula-I) is compared with the Polar distance formula (Distance Formula-II).

The distance between two n-dimensional Polar coordinates is basically specified with arc length and angles among two Polar coordinates from origin. We have used the distance formula described in [5] to find the distance between two n-dimensional Polar coordinates. The quality of clustering is evaluated using Silhouette internal evaluation measure as it combines the idea of both cohesion and separation of clusters.

In order to obtain better cluster quality, we implemented the concept of iterative clustering using Silhouette index. Based on the previous clusters, Silhouette score input parameter for next iteration is decided. We used the input parameter automation concept to achieve best partition of data set.

Section 2 describes the proposed clustering algorithm for Polar coordinates. Evaluation metric used for comparison is described in Sect. 3. Subsequent sections present experimental result and observations. Successively we provide the conclusion in Sect. 6.

2 Modified Density-Based Clustering Approach for Polar Coordinates

In order to design iterative clustering algorithm, we have proposed modification in our previously proposed algorithm [1]. The main drawback of most of the partition based algorithm is the requirement to provide the expected number of clusters or any other parameter. In this advanced version, we proposed an iterative method that works on quality metrics and accordingly obtained the results. We have used Silhouette internal evaluation measure to assess the cluster quality. Result of two successive iterations (Silhouette score) are compared. Based on comparison, input parameter of next iteration is modified. After performing certain iterations, we stop execution of the program and maximum Silhouette score is selected as result and accordingly the respective number of clusters are assumed to be good partitions for the given data set. The error parameter (D) is initialized and during iterative task it gets modified based on Silhouette score comparison of previous and current iteration. The distance between two n-dimensional Polar coordinates (Distance Formula-II) is computed using (1) and (2) [6].

$$\begin{aligned} d\left(x, x'\right) &= R\gamma = R\ \cos^{-1}\left(\left(X, X'\right) / (X, X)\left(X', X'\right)\right) \\ &= R\ \cos^{-1}\left(\mathrm{COS}\,\theta\,\mathrm{COS}\,\theta' + \sin\ \theta\ \sin\ \theta'\ \cos\ \gamma\right) \end{aligned} \tag{1}$$

where separation angle gamma is given as follows:

$$\mathrm{Cos} = \cos\left(\phi - \phi'\right) \prod_{i=1}^{d-2} \sin\ \theta\, i\ \sin\ \theta^{`} i + \sum_{i=1}^{d-2} \cos\ \theta\, i\ \cos\ \theta' i \prod_{i=1}^{d-2} \sin\ \theta j\ \sin\ \theta' j \tag{2}$$

Algorithm 1 describes the steps of modified density-based algorithm for Polar coordinates. Initially, map all data point of data set represented in Cartesian coordinate to Polar coordinate system, then the following steps starts.

Input:

$S =$ sl, $s2$,……, sn…… Set of n data items represented using Polar coordinates.

Si.di represents the ith dimension of the ith object [1].

minDistance (Minimum allowable difference between arc length and angles among two n-dimensional Polar coordinates from origin to be in one cluster).

Output:

Well separated and compact group of objects.

Algorithm 1 Modified Density-Based Polar Clustering Approach

```
1.  Initialize the input parameter minDist by adding Error parameter ($\Delta$).
2.  Repeat
3.  For each m-dimensional data point of data set, calculate dissimilarity of current data point with remaining data points of data set using Polar distance formula (Distance Formula-II)
4.  Determine the connectivity of data point using density-based function Connected(i) as follows:
    Connected(i) = {Sj} such that,
      i.  For all dimensions dp (1<=p<=m)
      ii. Dist(Si.dp, Sj.dp) < minDistance
5.  if data points agree upon Connected(i) function then they goes in one cluster then
6.  Mark si and all s j as clustered
7.  end if
8.  Until all data point are not clustered.
9.  Evaluate the quality of cluster of each iteration using Silhouette internal measure
10. Compare Sil Score of two successive iterations and update input parameter in following manner.
11. If(Prev_Sil>New_Sil) then
12. $\Delta = \text{constant value} * \Delta$
13. Else
14. $\Delta = -\text{constant value} * \Delta$
15. End if
16. Go to step 3.
17. Select max Silhouette score as result and stop.
```

3 Evaluation Measure

Evaluation measure is used to evaluate the goodness of cluster. We have used the Silhouette evaluation measure which combines both cohesion and separation. Silhouette index is the widely used internal cluster validation measure for crisp clustering. It evaluates the goodness of cluster structure without internal information. Silhouette index is the measure of how similar an object to its assigned cluster compared to other clusters. Its value ranges from −1 to +1. Negative value indicates an object has been wrongly assigned to the cluster. It could be member of other cluster. Value 0 indicates an object is on the verge of two natural clusters. Positive high value indicates an object is well clustered. Silhouette index can be obtained using the following formula (3).

$$S(i) = b(i) - a(i)/\max\{a(i), b(i)\} \tag{3}$$

Table 1 Data sets description

Data set name	No. of data points	No. of dimension	No. of clusters
Synthetic data set-I	65	2	3
Standard Iris data set	150	3	3
NSL-KDD	1000	41	2
Letter recognition	6111	16	8

where

$a(i)$ is the average dissimilarity of i with all other objects within the same cluster.
$b(i)$ is the average dissimilarity of i to the remaining clusters

4 Experiments

In this paper, we presented the observations of clustering with Polar coordinate system. We applied the proposed clustering approach on Cartesian mapped Polar coordinates using two different distances formula. The performance of both the distance metrics are compared using Silhouette internal evaluation measure. For experiment, we used a standard as well as a real data set of different dimensions (Table 1).

For experimental purpose, we used four different data sets of different dimensions. In this paper, the synthetic data set of 65 objects is used which is developed by Lingras et al. [6]. Data set comprise of clearly separable three clusters and five ambiguous objects.

Second, we used a standard Iris Data Set obtained from UCI machine learning repository [7]. The four-dimensional standard Iris data set consist of three distinct classes, namely Setosa, Vesicolour, and Virginica of 50 instances each.

Third data set we used for experiment is NSL-KDD intrusion detection data set with the aim to cluster each record into two groups, i.e., normal and attacks. NSL-KDD data set contains 41 attributes. Feature selection method is applied to reduce the dimensionality of data set which enhances the accuracy of clustering algorithm. We used Weka data mining tool to apply a Correlation attribute evaluation method. After applying feature selection method, three most significant numeric attributes were selected out of 41 attributes which are listed in Table 2. In our experiment, we used only 1000 instances.

Next to check the scalability of our proposed algorithm, we have used the data set of large size. We obtained Letter Recognition [8] data from the University of California Irvine machine learning repository. Data set consist of 20,000 records of A to Z capital letters of 20 different fonts. Instead of using all 26 characters, we decide to work with eight distinct characters A, H, L, M, O, P, S, Z, and 6111 records.

Table 2 Three significant attributes obtained by a Correlation attribute evaluation method rank feature set

Rank	Rank feature set
0.749	same_srv_rate
0.719	dst_host_srv_count
0.692	dst_host_same_srv_rate

Table 3 Cluster evaluation result for the synthetic data set-I

Connected approach	No. of cluster	Best silhouette score
DistFormula-I	3	0.6091
DistFormula-II	8	0.7817

Table 4 Cluster evaluation result for the standard Iris data set-II

Connected approach	No. of cluster	Best silhouette score
DistFormula-I	3	0.4506
DistFormula-II	2	1.0

In this paper, our proposed algorithm for Polar coordinates using Euclidean distance [1] is compared with the same using Polar coordinates distance formula.

5 Observations

This paper presents the comparison between our earlier proposed algorithm [1] using Euclidean distance and its modified distance, i.e., Polar distance formula with the experimental results.

Table 3 shows the experimental results of Polar coordinates clustering for synthetic data set-I. We observed that clustering results obtained using DistFormula-II (Polar coordinates distance formula) partitioned the 65 objects into 8 clusters. Out of eight clusters, three clusters show the correct partition and remaining five clusters contains ambiguous objects separately. Whereas, DistFormula-I simply partitioned the 65 objects into 3 clusters. No separate cluster is generated for ambiguous objects. So, we conclude that DistFormula-II (Polar distance) not only explicitly identifies ambiguous objects but also generates compact and well-separated clusters.

On the other hand, in Table 4 for standard Iris data set, we observed that the proposed algorithm using DistFormula-II partitioned 150 objects into 2 clusters instead of 3 clusters (i.e., exact number of clusters) still it yields separate and compact cluster compared to that of exact number of clusters generated using DistFormula-I.

In the experimental result of Table 5, we observed a major difference in Silhouette score of clusters obtained using DistFormula-I and DistFormula-II. The negative Silhouette score indicates possibly wrong assignments of instances. For the Letter

Table 5 Cluster evaluation result for the Letter recognition set-III

Connected approach	No. of cluster	Best silhouette score
DistFormula-I	34	−0.7099
DistFormula-II	8	0.5150

Table 6 Cluster evaluation result for the NSL-KDD data set-IV

Connected approach	No. of cluster	Best silhouette score
DistFormula-I	5	0.2192
DistFormula-II	2	0.99

recognition, data set of 6111 instances DistFormula-II yields exact number of well-separated clusters. Table 6 shows the experimental result obtained for NSL-KDD data set. The proposed algorithm using DistFormula-II partitioned the 1000 instances represented in Polar coordinates into 2 well separated and compact clusters. Whereas, same algorithm using DistFormula-I generates five clusters for the same instances, which are not compact as well as well separate as compared to the clusters generated using DistFormula-II.

6 Conclusion

We explore the possibilities of clustering with Polar coordinates using two different distance formulas. The performance of the clustering obtained using two different distance formulas are compared using Silhouette internal cluster evaluation measure. From various experiments, we observed that data points represented using Polar coordinate system can be better clustered using Polar distance formula than Euclidean distance formula.

References

1. Joshi, M.R., Patil, Y.S.: Analysis of change in coordinate system on clustering. In: 2016 IEEE Conference on Current Trends in Advanced Commuting (ICCTAC), pp. 1–7, March 2016
2. Ding, Shifei, Fulin, Wu, Qian, Jun, Jia, Hongjie, Jin, Fengxiang: Research on data stream clustering algorithms. Artif. Intell. Rev. **43**(4), 593 (2015)
3. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Surv. (CSUR) **31**(3), 264–323 (1999)
4. He, J., Tan, A.-H., Tan, C.-L., Sung, S.-Y.: On quantitative evaluation of clustering systems. In Clustering and Information Retrieval, vol. 11 of Network Theory and Applications, pp. 105–133. Springer, US (2004)
5. Howard S., COHL: Fundamental Solution of Laplace's Equation in Hyperspherical Geometry. In: Symmetry, Integrability and Geometry: Methods and Applications, SIGMA 7, 108, 14 p. (2011)

6. Lingras, P., Chen, M., Miao, D.: Rough multi-category decision theoretic framework. In: Wang, G., Li, T., Grzymala-Busse, J., Miao, D., Skowron, A., Yao, Y. (eds.) Rough Sets and Knowledge Technology, vol. 5009 of Lecture Notes in Computer Science. Berlin/Heidelberg, pp. 676–683 (2008)
7. Frank, A., Asuncion, A.: UCI Machine Learning Repository (2010)
8. Newman, D.J, Asuncion, A.: UCI Machine Learning Repository. University of California, Irvine, CA (2007). http://www.ics.uci.edu/~mlearn/MLRepository.html

Dynamic Modelling of Runoff in a Watershed Using Artificial Neural Network

Sandeep Samantaray and Dillip K. Ghose

Abstract In this work, an attempt has been made to measure the correlation between precipitation and infiltration loss and modelling are developed using Recurrent Neural network (RNN), and Radial Basis Neural Network (RBNN), to predict runoff. Models are evaluated using the statistical parameters mean square error (MSE) training, MSE testing, Root Mean Square Error (RMSE) training, RMSE testing and coefficient of determination (R^2). The output of this work will suggest the planning, design and management of water bound structures for developing the watershed. The performance of this work is compared and mapped with R^2 value. Results suggest that estimation of runoff is convenient to RNN as compared to RBNN. Both RNN and RBNN are better performing in complex data sets of the proposed watershed.

1 Introduction

In water resources management, the judgment of the accessibility of runoff potential is essential. Computation of runoff from precipitation and infiltration loss along with atmospheric parameters is required for any watershed. Runoff measurement in a watershed is dependent upon the period of availability of rainfall and runoff. Precipitation infiltration loss atmospheric parameter is co-related to each other. Such co-relations are used to measure the runoff from the observed precipitation, temperature and infiltration loss. Precipitation is the discharge of water from the atmosphere to the earth. Key input of water to a catchment area depends on hydrological study. In this framework, this study is an attempt for monitoring the rainfall infiltration loss. Runoff from the watershed is the flow of water from different surface areas to an outlet. Complex relationships between rainfall and infiltration loss are to understand through planning, design and management of watershed. This is obtained through hydrological modelling.

S. Samantaray · D. K. Ghose (✉)
Department of Civil Engineering, National Institute of Technology, Silchar 788010, Assam, India
e-mail: dillipghose2002@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_59

Kothyari [6] formulated a method for predicting monthly runoff during monsoon. Carriere [1] expanded runoff hydrograph system with integration of recurrent back-propagation neural network. The cascade correlation algorithm is used to predict two-year peak discharge from watersheds in continental United States by Muttiah et al. [8]. The radial basis function neural networks have similar modelling results to that of the multilayer perceptron neural network for rainfall runoff model [2, 9]. The advances of hybrid neural networks with conceptual model have proved considerable attention [5, 7]. El-Shafie and Noureldin [3] developed generalized neural network model to prevail over conventional forecasting techniques. ANN approach is used for forecasting of long-term reservoir inflow with monthly available data [4]. Literature studies indicate ANNs are useful to predict runoff. The objective of the study is to quantify runoff from rainfall and infiltration loss in a complex system.

2 Study Area and Data Collection

Loisinga watershed of Bolangir district, Odisha, India, is taken into consideration for the proposed study area. The study is made for predicating runoff in three watersheds to assess the drainage capacity of watershed during monsoon period ranging from 1993 to 2012. The watersheds are located in the upper part of Hirakund reservoir. The latitude 20°52′0″N and longitude 83°31′0″E are the geocoordinate of the watersheds shown in Fig. 1. It is located at an elevation of 162 m from mean sea level (MSL).

Daily average precipitation, daily average temperature for monsoon month (May to October) from the period 1993–2016 spanning over 24 year are collected from IMD Bhubaneswar. Daily runoff and daily infiltration rate data are collected from soil conservation office Bolangir.

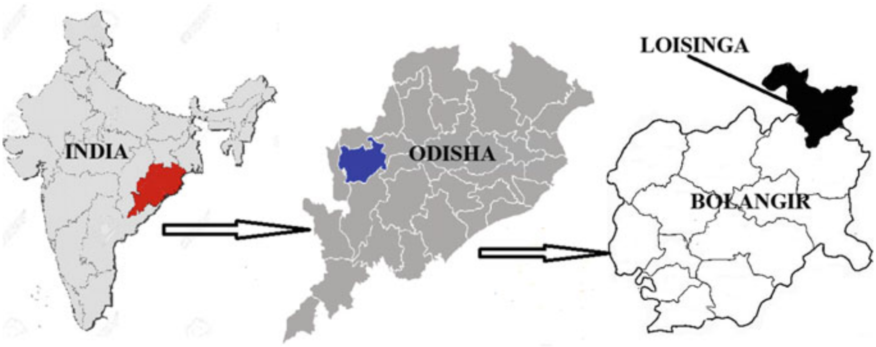


Fig. 1 Study area: Loisinga

3 Methodology

3.1 Radial Basis Neural Network

RBNN is a type of feed-forward neural network for applying the algorithms to supervised learning. Basically, RBNN is composed of large number of interconnected artificial neurons and is computed through interaction of input layer, hidden layer and output layer as represented in Fig. 2.

3.2 Recurrent Neural Network

RNN is a modification over BPNN and RBNN through addition of context unit. RNN consists of four layers, an input layer, a hidden layer, a context layer and an output layer. Each input unit is connected to every hidden unit through appropriate weight and bias. Context unit is a transfer integration of weights directed from input and hidden layer. The downward connections allow the context units to store the outputs of the hidden nodes at each time step; and distributed upward links feed them back as additional inputs. Here, the recurrent connections permit the hidden units to recycle the information over multiple steps, through temporal information with sequential input and the target function (Fig. 3).

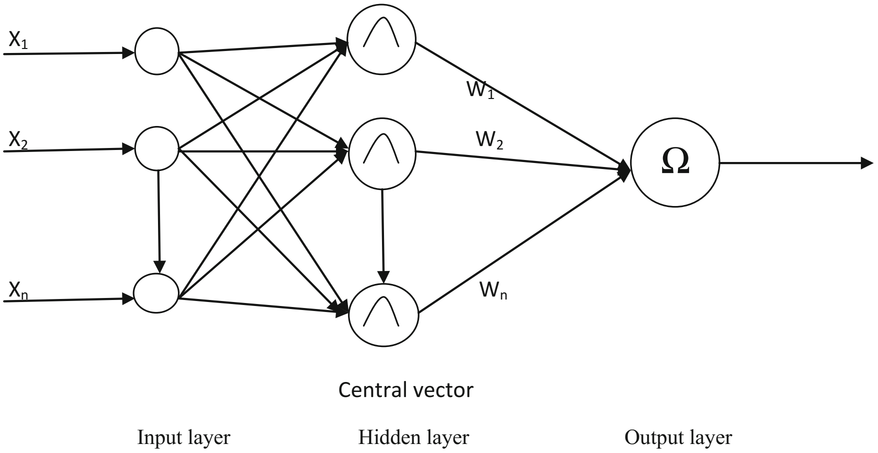


Fig. 2 Architecture of radial basis neural network

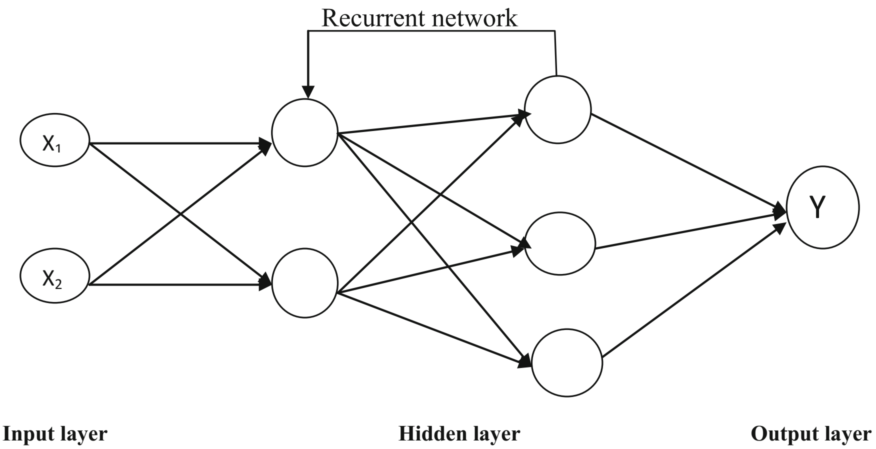


Fig. 3 Architecture of recurrent neural network

Table 1 Results for RBNN

Model input	Architecture (spread value)	MSE		RMSE		R^2	
		Training	Testing	Training	Testing	Training	Testing
Precipitation, average temperature, infiltration loss	3-0.2-1	0.00681	0.00542	0.00733	0.00487	0.8984	0.8731
	3-0.3-1	0.00331	0.00659	0.00575	0.00257	0.9052	0.8857
	3-0.5-1	0.00344	0.00498	0.00587	0.00223	0.8731	0.8465
	3-0.7-1	0.00363	0.00791	0.00602	0.00245	0.8899	0.8610
	3-0.9-1	**0.00281**	**0.00642**	**0.00531**	**0.00377**	**0.9233**	**0.9029**

Bold indicates the best model architecture for different membership function

4 Results and Discussions

The results of Radial basis neural networks are presented in Table 1 with various spread values are taken for simulation. Here spread values are considered within range of 0–1, i.e. preferably 0.2, 0.3, 0.5, 0.7 and 0.9 for predicating runoff from the considerable input parameters for mapping output. It is found that with a spread value 0.9 the RBNN shows best performance with architecture having 3-0.9-1 which possess MSE training 0.00281 testing 0.00642, RMSE training 0.00531 testing 0.00377 and coefficient of determination training 0.9233 testing 0.9029.

The layer recurrent neural network the results are discussed below for Loisinga station. For Tan-sig, Log-sig, and Purelin transfer function with 3-3-1, 3-5-1, 3-7-1, 3-8-1 and 3-9-1 architectures are taken into consideration for computation of performance. For Tan-sig function the best model architecture is found to be 3-7-1 which possess MSE training architecture MSE training value 0.000265, MSE testing value 0.001722, RMSE training value 0.001275, RMSE testing value 0.001099 and coefficient of determination value training 0.9578, testing value 0.9351. Similarly for Log-sig function, the best model architecture is found to be 3-5-1 which possess MSE training value 0.000453, MSE testing value 0.001887, RMSE training value 0.002492, RMSE testing value 0.004343 and coefficient of determination value training 0.9416, testing value 0.9265. Also same for Purelin function, the best model architecture is found to be 3-5-1 which possess MSE training value 0.000500, MSE testing value 0.003038, RMSE training value 0.002643, RMSE testing value 0.005512 and coefficient of determination value training 0.9179, testing value 0.9019. Detailed results for other transfer functions are tabulated in Table 2.

5 Simulation

The graphs with best values for runoff from precipitation, average temperature and infiltration loss using Recurrent Neural Network and Radial Basis Neural Network with Tan-sig and Log-sig transfer function are presented below at Loisingha. The best value for each evaluating criteria is represented. The best values results in the variation between the observed and predicted runoff are shown in graph (Figs. 4 and 5).

5.1 *Assessment of Actual Runoff Versus Simulated Runoff at Loisinga During Testing Phase*

The variation of actual runoff versus simulated or predicted runoff is showing in Fig. 6. Results show that the estimated peak runoff is 348.3953 mm, and 335.8461 mm

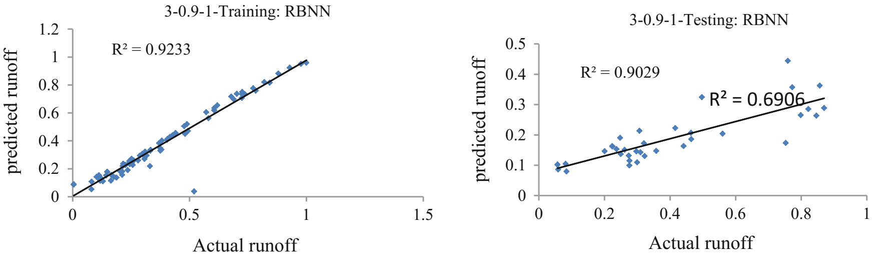


Fig. 4 Best actual versus predicted runoff model using RBNN

Table 2 Results for RNN

Model Input	Transfer Function	Architecture	MSE		RMSE		R^2	
			Training	Testing	Training	Testing	Training	Testing
Precipitation, average temperature, infiltration loss	Tan-sig	3-3-1	0.000601	0.005558	0.009330	0.009251	0.9112	0.8907
		3-5-1	0.000725	0.006006	0.001797	0.004480	0.9354	0.9102
		3-7-1	**0.000265**	**0.001772**	**0.001275**	**0.001099**	**0.9578**	**0.9351**
		3-8-1	0.000503	0.004391	0.003207	0.006196	0.8947	0.8704
		3-9-1	0.000718	0.006215	0.003439	0.006492	0.9018	0.8867
	Log-sig	3-3-1	0.000604	0.006200	0.003232	0.007875	0.9038	0.8887
		3-5-1	**0.000453**	**0.001887**	**0.002492**	**0.004343**	**0.9416**	**0.9265**
		3-7-1	0.000700	0.003831	0.002643	0.004277	0.9205	0.9046
		3-8-1	0.000639	0.002719	0.003735	0.00214	0.9185	0.8980
		3-9-1	0.000504	0.002994	0.00325	0.005472	0.9328	0.9122
	Purelin	3-3-1	0.000981	0.003645	0.003130	0.006037	0.9031	0.8995
		3-5-1	**0.000500**	**0.003038**	**0.002643**	**0.005512**	**0.9179**	**0.9019**
		3-7-1	0.000616	0.007496	0.003411	0.008659	0.9017	0.8900
		3-8-1	0.000703	0.004487	0.003165	0.006699	0.8934	0.8719
		3-9-1	0.000827	0.007039	0.003563	0.008392	0.8805	0.8622

Bold indicates the best model architecture for different membership function

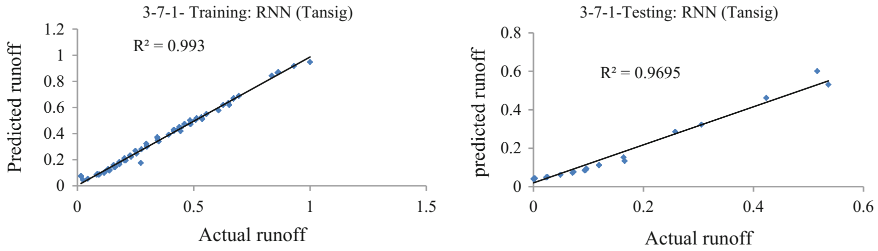


Fig. 5 Best actual versus predicted runoff model using RNN

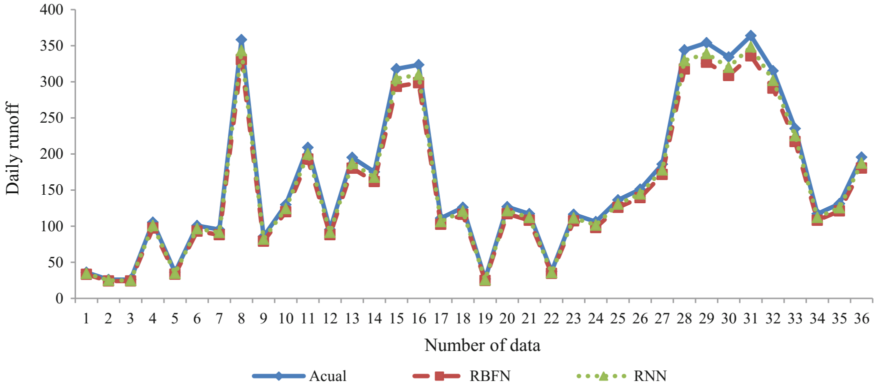


Fig. 6 Actual versus simulated runoff using RNN and RBFN at Loisinga in testing phase

for RRN and RBNN against the actual peak 363.7454 mm and estimated minimum runoffs are 25.9087, 23.9215 and 24.8135 as actual runoff and predicted runoff by RBNN and RNN, respectively, for the watershed Loisinga.

6 Conclusions

In this study daily precipitation, infiltration loss and temperature data have been considered for predicting runoff. Two neural networks are used to estimate daily runoff during monsoon period with the evaluation criteria MSE, RMSE and R^2 value. RNN performs best, with architecture 3-7-1 following Tan-sig transfer function and 3-5-1 with Log-sig transfer function. RBNN performs best with architecture 3-0.9-1 with the spread value 0.9. Taken as a whole the results suggest that RNN performs better than RBNN with a range of confidence limit of performance around ±10%. This work will help for conservation of water, controlling soil erosion and designing hydraulics structure for development of the watershed. The results may be improved by an integral approach of combined technique which is to be investigated for improving the models scenarios in future scope.

References

1. Carriere, P., Mohaghegh, S., Gaskari, R.: Performance of a virtual runoff hydrograph system. J. Water Resour. Plann. Manage. **122**(6), 421–427 (1996)
2. Dawson, C.W., Wilby, R.L.: Hydrological modelling using artificial neural networks. Prog. Phys. Geogr. **25**(1), 80–108 (2001). https://doi.org/10.1177/030913330102500104
3. El-Shafie, A., Noureldin, A.: Generalized versus non-generalized neural network model for multi-lead inflow forecasting at Aswan High Dam. Hydrol. Earth Syst. Sci. Discuss. **7**(5), 7957–7993 (2010)
4. Faridah, O., Mahid, N.: Reservoir inflow forecasting using artificial neural network. Int. J. Phys. Sci. **6**(3), 434–440 (2011)
5. Jain, A., Srinivasulu, S.: Integrated approach to model decomposed flow hydrograph using artificial neural network and conceptual techniques. J. Hydrol. **317**(3–4), 291–306 (2006). https://doi.org/10.1016/j.jhydrol.2005.05.022
6. Kothyari, U.C.: Estimation of monthly runoff from small catchment in India. Hydrol. Sci. J. **40**(4), 533–541 (1995)
7. Lee, D.S., Jeon, C.O., Park, J.M., Chang, K.S.: Hybrid neural network modeling of a full-scale industrial wastewater treatment process. Biotechnol. Bioeng. **78**(6), 670–682 (2002). https://doi.org/10.1002/bit.10247
8. Muttiah, R.S., Srinivasan, R., Allen, P.M.: Prediction of two-year peak stream discharges using neural networks. J. Am. Water Resour. Assoc. **33**(3), 625–630 (1997)
9. Zakermoshfegh, M., Ghodsian, M., Salehi Neishabouri, S.A.A., Shakiba, M.: River flow forecasting using neural networks and auto-calibrated NAM model with shuffled complex evolution. J. Appl. Sci. **8**, 1487–1494 (2008). https://doi.org/10.3923/jas.2008.1487.1494

Estimation of Aquifer Potential Using BPNN, RBFN, RNN, and ANFIS

Umesh Kumar Das, Sandeep Samantaray, Dillip K. Ghose and Parthajit Roy

Abstract In this paper, aquifer potential in terms of sensitivity of a well in arid region is measured. Four techniques such as Backpropagation neural network (BPNN), Radial basis function network (RBFN), Recurrent neural network (RNN), and Adaptive Neuro-Fuzzy Inference System (ANFIS) are used to predict the aquifer potential of the well. Measured specific drawdown of the well is considered as input and evaluated aquifer loss coefficient as output for developing the efficiency of model. The complexity of aquifer characteristics of the region is measured through the sensitivity of the developed models. Results of all techniques explain the variation of aquifer characteristics in arid region. Among all proposed models, ANFIS executes best for mapping the sensitivity of aquifer. Overall results show the integrity of performances to understand the complex behavior of aquifer in decreasing order of ANFIS, RNN, BPNN, and RBFN.

1 Introduction

Aquifer loss occurs due to head loss during flow of water to the well screen. During flow process, water passes rapidly through the well screen, and migrated to the aquifer immediately next to screen. In a step-drawdown test, water is pumped at a known value of discharge; water levels and time are recorded until drawdown begins to stabilize.

The drawdown in a pumping well comprises aquifer loss and well loss [2]. Both the components of drawdown are obtained by using the drawdown observed during a step-drawdown test. If a well is pumped and tested at various constant discharge rates until the drawdown stabilizes, then it is known as step-drawdown test. Generally, three or preferably more steps of pumping rate are considered for determination of aquifer loss coefficient. Step-drawdown test measures the performance of a well that can be used to determine well efficiency and also evaluate an optimal pumping rate

U. K. Das · S. Samantaray · D. K. Ghose (✉) · P. Roy
Department of Civil Engineering, National Institute of Technology, Silchar, Assam, India
e-mail: dillipghose2002@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_60

for the well. Water levels in a pumping well decrease with pumping period as well as with increased pumping rate. This fall in the water level is known as drawdown.

Graphical methods for the determination of components of drawdown were proposed by Jacob [2], Rorabaugh [5] and Sheahan [7]. Optimization method for determination of well loss was proposed by Singh [8]. An efficient, stable ANN model was proposed for predicting groundwater level in south-east Punjab, India by Lohani and Krishan [3]. Nayak et al. [4] recommended input sensitivity analysis with the exclusion of antecedent values of the water level time series to capture the recharge time for the aquifer. Sethi et al. [6] suggested that ANN used to predict water table depth in a hard rock aquifer with reasonable accuracy for short-term data. Adamowski and Chan [1] developed WA-ANN model to provide more accurate monthly average groundwater level prediction as compared to ANN and ARIMA model.

In this paper, an approach is presented which uses all the observed drawdown in a step-drawdown test to determine aquifer loss coefficient. A BPNN, RBFN. RNN, and ANFIS network is trained using all the measured drawdown and their corresponding pumping rates to determine aquifer loss coefficient. The networks are trained using specific drawdown as input and corresponding aquifer loss coefficient as output. The trained model gives aquifer loss coefficient as output when presented with specific drawdown as input. The result obtained using all the abovementioned approaches are in a good agreement with each other.

2 Study Area

Mahulpali village is located in Kisinda Tehsil of Sambalpur district in Odisha, India shown in Fig. 1. Mahulpali is surrounded by Sambalpur Tehsil towards North, Dhankauda Tehsil towards North, Jujomura Tehsil towards East, Burla Tehsil towards west and its coordinate is 21.8733°N, 84.4304°E. Data samples are observed by water level reorder to measure the drawdown of the well at Mahulpali.

3 Methodology

Artificial neural network (ANN) is a parallel processing tool for modeling. An ANN consists of input layer, hidden layer and output layer. Input layer neurons transmit information to the neurons in the hidden layer and passing over to output layer. The training of the network is done by exposing the network to a set of input data. The information is then processed by the hidden layer neurons. Finally, the network output is given by output layer neurons.

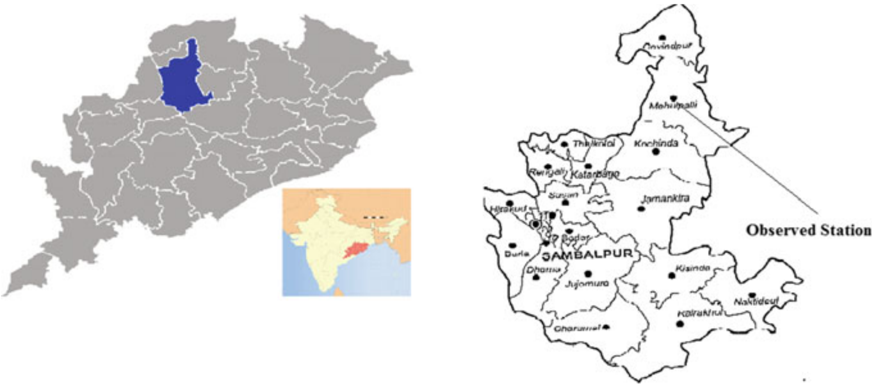


Fig. 1 Command area map of showing observing well

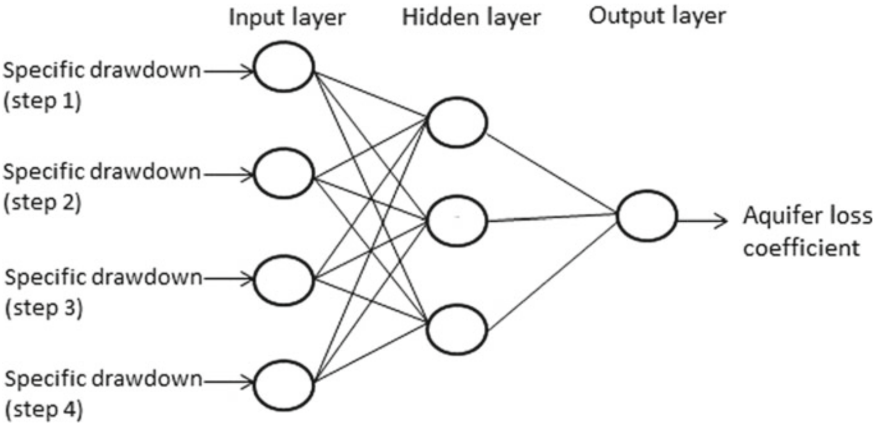

Input layer
Hidden layer
Output layer
Specific drawdown (step 1)
Specific drawdown (step 2)
Specific drawdown (step 3)
Specific drawdown (step 4)
Aquifer loss coefficient

Fig. 2 Architecture of BPNN model for determination of aquifer loss coefficient

3.1 Back Propagation Neural Network Model

Backpropagation neural network has an input layer, one or more hidden layer and an output layer. Figure 2 represents the architecture of BPNN model. It is a synthetic method for training multilayer ANNs.

From the calculated net input with the proposed activation function, the outputs are calculated. Activation functions used in net are binary-sigmoidal function and bipolar-sigmoidal function.

If the weights are given as, $W = (w_{ij})$ in a matrix form, the net input to output unit y_j

$$y_j = \text{Net} = b + \sum_i x_{ij} w_{ij} \tag{1}$$

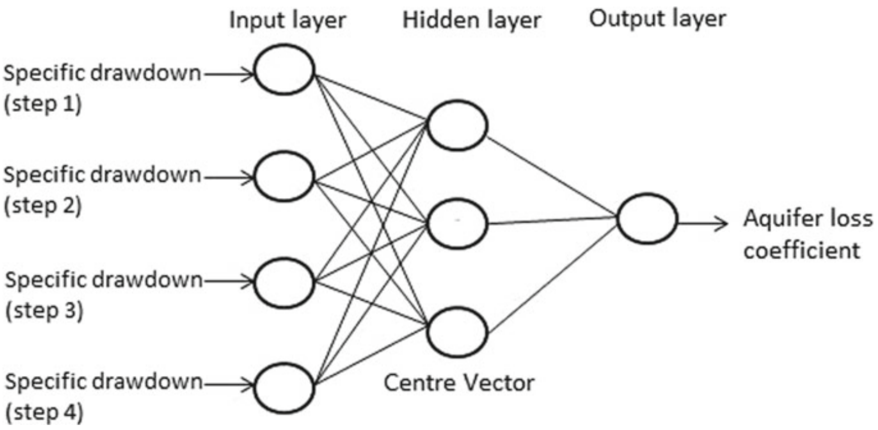

Input layer
Hidden layer
Output layer
Specific drawdown (step 1)
Specific drawdown (step 2)
Specific drawdown (step 3)
Specific drawdown (step 4)
Centre Vector
Aquifer loss coefficient

Fig. 3 Architecture of RBFN model for determination of aquifer loss coefficient

where b is bias

x_{ij} input from neuron i to j
w_{ij} weight of the neuron i to j

3.2 Radial Basis Function Network

A radial basis function neural network also has an input layer, a hidden layer, and an output layer. The neurons in the hidden layer comprises of Gaussian transfer functions whose outputs are inversely proportional to the distance from the center of the neuron shown in Fig. 3.

$$S(x) = \sum_{i=1}^{n} w_i v_i(x) \tag{2}$$

where $S(x)$ is the output and w_i, v_i are weights and biases, respectively.

$$\text{Euclidean distance} = \left\| I - V_j \right\| \tag{3}$$

3.3 Recurrent Neural Network

RNNs are a class of neural network where links between units forms a fixed cycle. This creates a state which allows the network to show dynamic behavior. Unlike

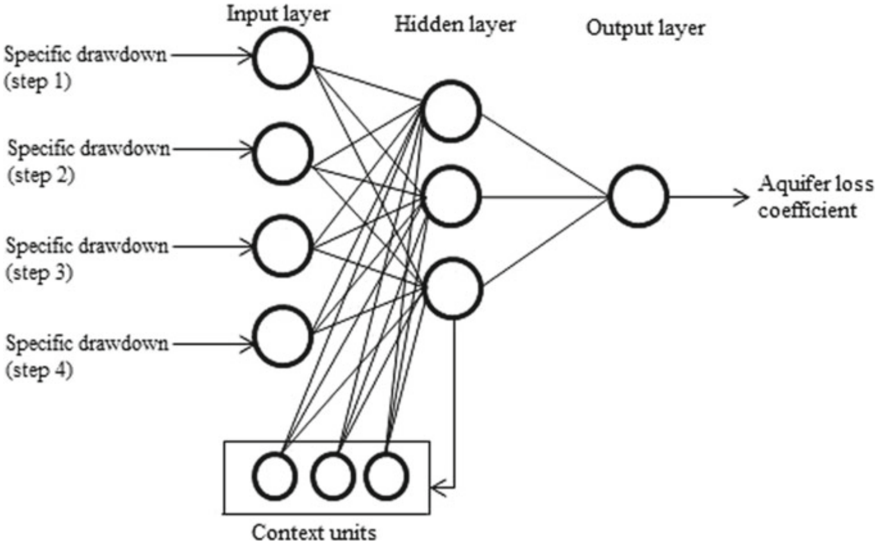

Input layer
Hidden layer
Output layer
Specific drawdown (step 1)
Specific drawdown (step 2)
Specific drawdown (step 3)
Specific drawdown (step 4)
Aquifer loss coefficient
Context units

Fig. 4 Architecture of RNN model for determination of aquifer loss coefficient

feed forward networks, RNN's uses their internal memory to process the past neural states (Fig. 4).

3.4 Adaptive Neuro-fuzzy Inference System

An adaptive neuro-fuzzy inference system is a type of ANN which is developed by Takagi-Sugeno fuzzy inference system as shown in Fig. 5. ANFIS integrates both neural networks and fuzzy principles. The advantage of ANFIS is that it captures the benefits of neural networks as well as the benefits of fuzzy logic. It uses a hybrid learning algorithm.

Training an ANFIS is done using grid partition so that Fuzzy Inference System (FIS) can be generated. Three numbers of gauss membership functions are used for each of the input patterns. Backpropagation optimization method is used in the training process. Training is done until the error is less than the specified error tolerance.

4 Results and Discussions

Measured drawdown of the well collected for 6 months on daily basis during non-monsoon period ranging from November to April 2016 for the bore well. Four sets

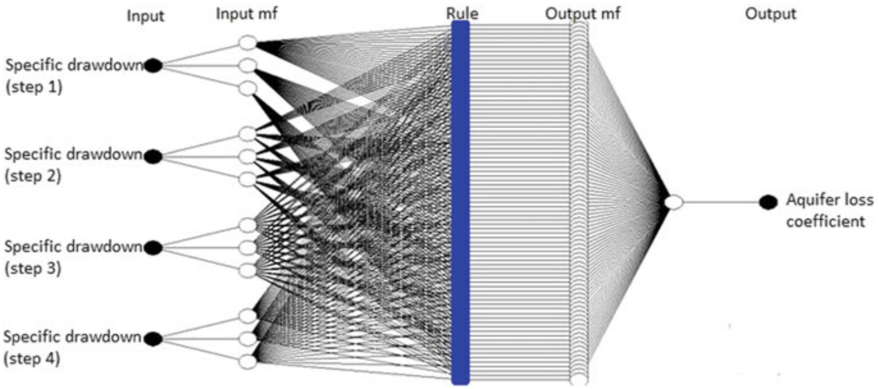

Input
Input mf
Rule
Output mf
Output
Specific drawdown (step 1)
Specific drawdown (step 2)
Specific drawdown (step 3)
Specific drawdown (step 4)
Aquifer loss coefficient

Fig. 5 Architecture of ANFIS for determination of aquifer loss coefficient

of specific drawdown measured 3 hours interval and is used as inputs for model development with four different techniques and aquifer loss coefficient as output. 70% of the data set is used for training and 30% of the data set is used for modeling the aquifer potential.

The performance of the four models, i.e., BPNN, RBFN, RNN, and ANFIS is assessed using Root Mean Square Error (RMSE), Coefficient of determination (R^2), and Mean Absolute Error (MAE).

$$\mathrm{RMSE} = \left[\frac{1}{N}\sum_{i=1}^{n}\left(Y_{\exp} - Y_{\mathrm{est}}\right)^2\right]^{1/2} \tag{6}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{n}\left|Y_{\mathrm{est}} - Y_{\exp}\right| \tag{7}$$

where $Y_{\exp}$ is the actual output or desired output and Y_{est} is the estimated output.

Results presented in Table 1 shows that 4-3-1 architecture using BPNN performs better as compare to other architecture performed by BPNN. The model efficiency for 4-3-1 architecture is 0.9584 during training and 0.9455 during testing. Similarly, Architecture 4-3-1 shows better with model efficiency 0.9774 in case of training phase and 0.9694 in case of testing phase for RNN techniques. Architecture 4-10-1 of RBFN performs better with model efficiency during training and testing is 0.9128 and 0.9032, respectively. Also in case of ANFIS, coefficient of determination is 0.9935 and 0.9789 during training and testing phase using Tri membership function. Among all the four techniques, ANFIS shows better as compared to RBFN, RNN, and BPNN techniques (Fig. 6).

Table 1 Performance of models for determination of aquifer potential

Techniques	Architecture /membership function	RMSE (days/m^2)		MAE (days/m^2)		Coefficient of determination (R^2)	
		Training	Testing	Training	Testing	Training	Testing
BPNN	**4-3-1**	**0.000299**	**0.002653**	**0.000212**	**0.002527**	**0.9584**	**0.9455**
	4-5-1	0.000281	0.002569	0.000193	0.002491	0.9461	0.9245
	4-10-1	0.000286	0.002611	0.000201	0.002510	0.8991	0.8651
RBFN	**4-10-1**	**0.000591**	**0.005212**	**0.000432**	**0.004513**	**0.9128**	**0.9032**
	4-15-1	0.000471	0.004825	0.000371	0.004183	0.8991	0.8816
	4-20-1	0.000528	0.005183	0.000418	0.004275	0.8848	0.8558
RNN	**4-3-1**	**0.000285**	**0.001209**	**0.000133**	**0.000795**	**0.9774**	**0.9694**
	4-4-1	0.000183	0.001096	0.000092	0.000638	0.9328	0.9294
	4-5-1	0.000259	0.001176	0.000117	0.000715	0.9214	0.9081
ANFIS	**Tri**	**0.000691**	**0.001425**	**0.000499**	**0.001120**	**0.9935**	**0.9789**
	Trap	0.000573	0.001329	0.000294	0.001028	0.9681	0.9513
	Gbell	0.000618	0.001384	0.000384	0.001095	0.9428	0.9338

Bold indicates best model architecture for determining aquifer potential

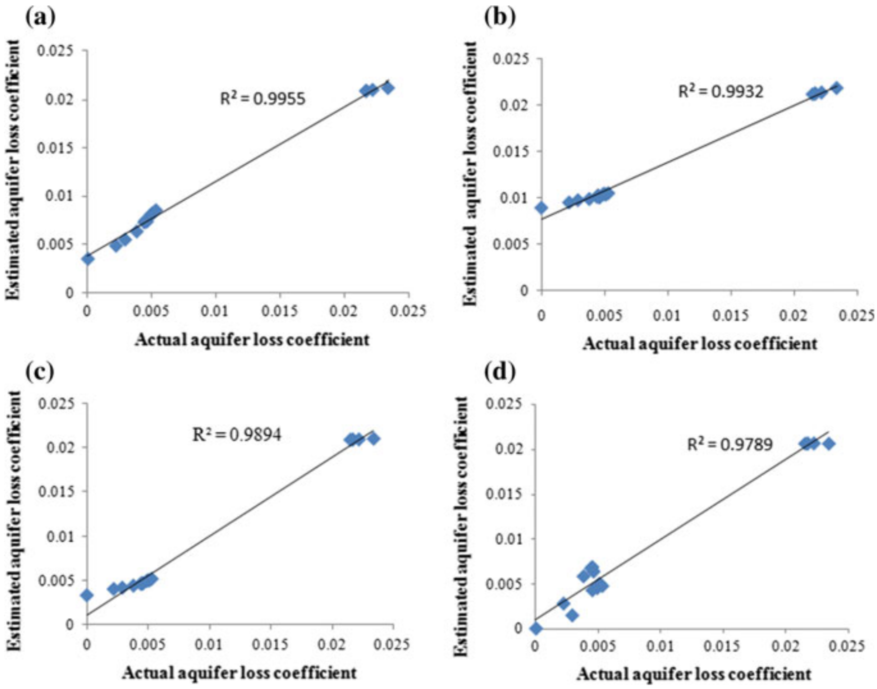


Fig. 6 Actual versus estimated aquifer loss coefficient during testing phase using **a** BPNN, **b** RBFN, **c** RNN, and **d** ANFIS

5 Conclusions

An advanced approach to determine the aquifer performance is developed using BPNN, RBFN, RNN, and ANFIS. The networks are trained to recognize input patterns of specific drawdown and output patterns of aquifer loss coefficient. The trained networks present aquifer loss coefficient as output when presented specific drawdown as input. The advantage of the present approach is the automated process of obtaining aquifer loss coefficient from specific drawdown data and the ability of the network to produce justifiable results.

References

1. Adamowski, J., Chan, H.F.: A wavelet neural network conjunction model for ground water level forecasting. J. Hydrol. **407**, 28–40 (2011). https://doi.org/10.1016/j.jhydrol.2011.06.013
2. Jacob, C.E.: Drawdown test to determine effective radius of artesian well. Trans. Am. Soc. Eng. **112**, 1047–1107 (1947)
3. Lohani, A.K., Krishan, G.: Application of artificial neural network for groundwater level simulation in Amritsar and Gurdaspur Districts of Punjab. India. J. Earth Sci. Clim. Change **6**(4), 1 (2015). https://doi.org/10.4172/2157-7617.1000274
4. Nayak, P.C., Rao, Y.R.S., Sudheer, K.P.: Groundwater level forecasting in a shallow aquifer using artificial neural network approach. Water Resour. Manage. **20**, 77–90 (2006). https://doi.org/10.1007/s11269-006-4007-zC
5. Rorabaugh, M.I.: Graphical and theoretical analysis of step-drawdown test of artesian wells. Proc. Am. Soc. Civ. Eng. Hyd. Div. **19**(1), 1–23 (1953)
6. Sethi, R.R., Kumar, A., Sharma, S.P., Verma, H.C.: Prediction of water table depth in a hard rock basin by using artificial neural network. Int. J. Water Resour. Environ. Eng. **2**(4), 95–102 (2010)
7. Sheahan, N.T.: Type-Curve Solution of Step-Drawdown Test. Ground Water **9**, 25–29 (1971)
8. Singh, S.K.: Well loss estimation: variable pumping replacing step step drawdown test. J. Hydraul. Eng. **128**(3), 343–348 (2002). https://doi.org/10.1061/(asce)0733-9429

Design and Implementation of a Scenario-Based Communication Model for VANETs in EXata

V. Ravi Ram, B. G. Premasudha and R. Suma

Abstract Vehicular Ad Hoc Network (VANET) is a promising area of research and development as it has remarkable role in improving safety of vehicles on road, efficient traffic management, and providing comfort to commuters. The typical characteristics of VANET environment pose several challenges for VANET application developers. There is always a demand for scalable communication mechanisms suitable for low- and high-speed vehicles, information dissemination techniques that suit sparse and dense traffic conditions. Thus, the reliability of any VANET application strongly depends upon the robustness of the underlying communication architecture. Any VANET communication model shall strictly adhere to the wireless access standards and specifications laid down for vehicular communications. In this work, we have designed and implemented a basic VANET communication model that facilitates both Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) communication adhering to IEEE 802.11p standards. The Physical and Link Layer extensions of WAVE standard have been integrated to EXata5.1 (network emulator) to support IEEE 802.11p communication. Necessary Graphical User Interface and library level extensions were made to create VANET scenarios. To demonstrate and evaluate our communication model, urban scenarios with different node densities were created in EXata. The communication scenarios were emulated with predefined mobility of nodes, CBR as the traffic application and Bellman Ford as the default routing protocol. Using EXata analyzer, the performance metrics were analyzed and found that our basic VANET communication model offers good throughput at high vehicle densities with a soft delay constraint.

V. Ravi Ram (✉) · R. Suma
Department of MCA, SSIT, Tumkur, Karnataka, India
e-mail: raviramv@gmail.com

R. Suma
e-mail: sumaraviram@gmail.com

B. G. Premasudha
Department of MCA, SIT, Tumkur, Karnataka, India
e-mail: bgpremasudha@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_61

1 Introduction

Vehicular Ad hoc Network (VANET) is a subclass of Mobile Ad hoc Networks (MANETs) that enables Intelligent Transportation System (ITS) communication. VANET consists of vehicles with onboard communication devices, Global Positioning System (GPS), Road Side Units (RSUs) and any sensors if required to report the vehicle conditions. The vehicles in VANET communicate among themselves and also with the RSUs in their radio range. Each vehicle acts as a wireless router such that, the vehicles with in a range of 100–500 m can form an ad hoc network. VANETs support several wireless devices used in vehicles such as mobile phones, laptops, and Portable Digital Assistants [1]. There is a huge demand for robust VANET applications that support Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) Communication [2]. The active safety applications send messages relating to dangerous and abnormal road conditions and collision warnings. The public services include lane clearance for emergency vehicles and tracking of stolen vehicles. Applications for improved driving shall assist in enhanced route guidance. The infotainment applications include internet access, instant messaging, notifications on point of interest, automatic collection of toll and parking fee.

The typical network and environmental characteristics makes research in VANETs more challenging [3]. Any VANET communication model shall strictly adhere to the communication standards laid down for VANETs. These wireless access standards include communication protocols, security specifications, routing and addressing services, and interoperability protocols [4]. The details of VANET communication standards are discussed in Sect. 2. Based upon current, ongoing research in the ITS environment, we have identified a VANET Communication Evaluation Model [5], to identify the critical factors to be met by researchers engaged in ITS activities. Various design issues considered during the implementation of IEEE802.11p are discussed in Sect. 3. The integration of Physical and MAC layer DSRC extensions, GUI level modifications and enhancement to wireless libraries to support 802.11p communication are described in Sect. 4. Typical VANET urban scenarios that facilitate V2V and V2I communication were created using an urban terrain in EXata with varied number of nodes ranging from 10 to 60. The scenario-terrain, channel, node properties along with scenario screenshots are presented in Sect. 5. The scenario-based emulation process with tabulated results is described in Sect. 6. Finally, the concluding remarks on the performance of the model are presented in Sect. 7.

2 Wireless Access Standards for VANET Communication

Dedicated Short Range Communications (DSRC) standard facilitates both V2V and V2I communications to support several safety and non-safety VANET applications.

DSRC facilitates high data transmission rates with in a small communication range and is based on the IEEE 802.11a physical layer and 802.11 MAC layer

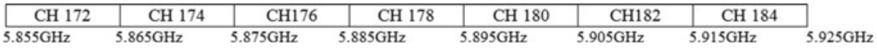


Fig. 1 Channels in DSRC band

[6]. DSRC facilitates half-duplex V2V communication using short range radio (300–1000 m) operating at high data rates ranging from 6 to 27 Mbps. The DSRC band (5.855–5.925 GHz) comprising of seven channels each of 10 MHz is shown in Fig. 1.

The DSRC band has two small zone service channels and two medium zone service channels designated for extended data transfer. Even though safety-related applications enjoy the highest priority across the channels, still two service channels are specially designated for safety related applications. In a typical VANET communication scenario, an RSU announces the availability of all applications to the On Board Units (OBUs) in its range, 10 times per second. Each OBU listening on channel 172, authenticates the RSU, executes safety applications and then switches channels to execute non-safety applications. Finally, the OBU returns to channel 172 and listens for the arrival of new safety applications.

Due to typical and dynamic characteristics of VANETs, conventional IEEE 802.11 Media Access Control (MAC) operations underperform when used in vehicular environment. The improvised version of DSRC is renamed as IEEE 802.11p Wireless Access in Vehicular Environments (WAVE) [7]. The top layers of IEEE 1609 standard handle the operational complexities of DSRC. The management activities specified in IEEE 1609.1 [8], the security protocols defined in IEEE 1609.2 [9], and the network-layer protocols defined in IEEE 1609.3 [10] regulates the functioning of WAVE applications. As IEEE 1609.4 [11] standard lies above IEEE 802.11p, it supports all higher layer operations without involving any of the physical channel access parameters. Any stationary WAVE device that hosts VANET application can act as a service provider. Similarly, any mobile device that runs the peer application can use those services. Figure 2 describes WAVE, IEEE 1609, IEEE 802.11p standards.

3 IEEE 802.11p Design Issues

The main challenges being faced by IEEE802.11p are frequency spectrum availability and fading. The system shall ensure low failure rate while executing safety related applications. Thus, 802.11p recommends few physical and MAC layer modifications to ensure robust and quick connection establishment among fast moving vehicles.

3.1 IEEE 802.11p Physical Layer

The 802.11p physical layer standard is similar to 802.11a standard and uses Orthogonal Frequency Division Multiplexing (OFDM) supporting eight data rates. The 802.11p standard defines 20, 10 and 5 MHz Physical layer modes by using half clocked mode and 10 MHz bandwidth. A comparison on several physical layer values is shown in Table 1.

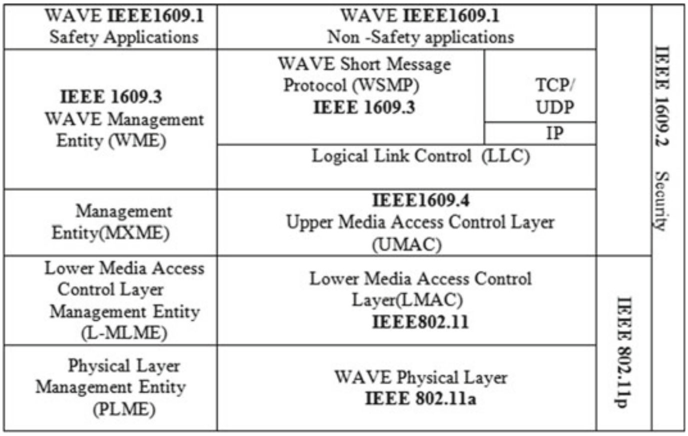

WAVE IEEE1609.1 Safety Applications
WAVE IEEE1609.1 Non -Safety applications
IEEE 1609.3 WAVE Management Entity (WME)
WAVE Short Message Protocol (WSMP) IEEE 1609.3
TCP/ UDP
IP
Logical Link Control (LLC)
Management Entity(MXME)
IEEE1609.4 Upper Media Access Control Layer (UMAC)
Lower Media Access Control Layer Management Entity (L-MLME)
Lower Media Access Control Layer(LMAC) IEEE802.11
Physical Layer Management Entity (PLME)
WAVE Physical Layer IEEE 802.11a
IEEE 1609.2 Security
IEEE 802.11p

Fig. 2 Wireless access standards for VANET [6]

Table 1 Comparison of the physical layer values of 802.11a and 802.11p

Parameters	IEEE 802.11p value(s)	Changes when compared to IEEE 802.11a
Bit rate (Mbit/s)	3, 4.5, 6, 9, 12, 18, 24, 27	Half
Modulation	BPSK, QPSK, 16QAM, 64QAM	Same
Code rate	1/2, 2/3, ¾	Same
No. of subcarriers	52	Same
Symbol duration	8 μs	Twice
Guard time	1.6 μs	Twice
FFT period	6.4 μs	Twice
Preamble duration	32 μs	Twice
Subcarrier spacing	0.15625 MHz	Half

Table 2 EDAC parameters for the control channel

Access category	CWmin	CWmax	AIFS
AC0(back ground)	CWmin	CWmax	9
AC1(best effort)	((CWmin+1)/2)-1	CWmin	6
AC2(video)	((CWmin+1)/4)-1	((CWmin+1)/2)-1	3
AC3(voice)	((CWmin+1)/2)-1	((CWmin+1)/2)-1	2

3.2 IEEE 802.11p MAC Layer

The IEEE 802.11p MAC layer is based on generic 802.11 MAC layer. The IEEE 802.11p uses CSMA/CA to minimize collisions and encourage unbiased channel access. Arbitrary Inter Frame Space (AIFS) and Back off are the two important parameters that characterize CSMA/CA operations. Any station intended to transmit data, shall sense the channel for availability AIFS, however if the channel is busy it has to execute Back off. IEEE 802.11p MAC layer depends upon on WAVE multichannel operation and 802.11e Enhanced Distribute Channel Access (EDCA). The EDCA facilitates four Access Categories (AC0–AC3) for each channel. Each AC has different contention parameters where AC3 has the top priority in accessing medium. The internal contention procedure among the ACs uses AIFS and Contention Window (CW) as the contention parameters. The CW defines the initial random back off time and its value depends upon CWmin which is the initial size of CW and after each failed transmission CW size is doubled till it reaches CWmax. Table 2 describes the EDAC parameters for the control channel.

3.3 General Properties

A basic communication model adhering to the 802.11p amendments made to 802.11standard has been designed with the general channel property/value pairs represented in Table 3.

4 Integration of 802.11p PHY and MAC Layer Extensions to EXata

The EXata [12] communications simulation platform (EXata) is a network emulator used to evaluate communication networks with high mobility of nodes accurately in a realistic manner. EXata makes use of software virtual network (SVN) to represent networks. As EXata supports hardware-in-the-loop capabilities, it facilitates interoperation of real time antennas and radios with SVN. Unless network simulators, an

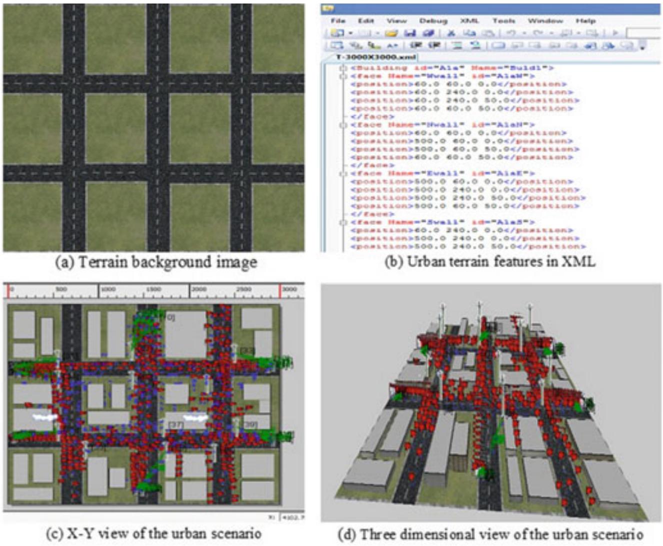
Table 3 Channel property/value pairs

General channel property	Value
No. of channels	2
Channels	PHY802.11pCCH, PHY802.11pSCH
Channel frequency	5.9 GHz
MAC protocol	MACDOT11p
Network protocol	IP
Path loss model	Two ray
Shadowing model	Constant
Signal propagation speed	3e8
Propagation limit (dBm)	−111.0
Propagation communication proximity	400

emulator mimics and behaves like a real network. Using EXata one can evaluate the performance of any network model or technology in a cost effective manner. Network emulation aids in making any design level changes and there by analyzing the impact of those changes on the performance of the system. EXata SVN is a cost effective alternative to physical test beds. EXata supports running real applications, analyzing tools, packet sniffing, etc. The QualNet extensions of the Physical and Link Layers of the WAVE 802.11p contributed by the authors in paper [13] have been integrated into EXata5.1 to support 802.11p. The C++ source file phy_802_11p.cpp and the header file phy_802_11p.h are used as the extensions at the physical layer. These extensions were used to initialize default values for data rate, transmission power, receiving sensitivity, etc. The C++ source file mac_wave_dot11p.cpp and the header file mac_wave_dot11p.h are used as the extensions at MAC layer. These extensions were used to manage the Sync Interval that coordinates multichannel access and to initialize EDCA parameters of 802.11P protocol. The PHY and MAC layer extensions have been successfully integrated with EXata to facilitate DSRC communication. The GUI level settings for the protocol models are made through phy_layer.prt and mac_layer.prt files. The source code for the wireless libraries supporting WAVE are made available through the C++ source files mac_dot11.cpp, mac_dot11-mgmt.cpp, phy_802_11.cpp with corresponding header files.

5 Creation of Application Oriented Scenarios

Typical VANET urban scenarios that facilitate V2V and V2I communication were created using an urban terrain in EXata with varied number of nodes ranging from 10 to 60. The image used as background for the urban terrain, a snapshot of the XML code used for the creation of the GUI elements such as buildings and the sample screen shots of the created scenario are shown in Fig. 3.



```
<Building id="A1a" Name="Buld1">
<face Name="Wwall" id="A1aW">
<position>60.0 60.0 0.0</position>
<position>60.0 240.0 0.0</position>
<position>60.0 240.0 50.0</position>
<position>60.0 60.0 50.0</position>
</face>
<face Name="Nwall" id="A1aN">
<position>60.0 60.0 0.0</position>
<position>500.0 60.0 0.0</position>
<position>500.0 60.0 50.0</position>
<position>60.0 60.0 50.0</position>
</face>
<face Name="Ewall" id="A1aE">
<position>500.0 60.0 0.0</position>
<position>500.0 240.0 0.0</position>
<position>500.0 240.0 50.0</position>
<position>500.0 60.0 50.0</position>
</face>
<face Name="Swall" id="A1aS">
<position>60.0 240.0 0.0</position>
<position>500.0 240.0 0.0</position>
<position>500.0 240.0 50.0</position>
```

(a) Terrain background image

(b) Urban terrain features in XML

(c) X-Y view of the urban scenario

(d) Three dimensional view of the urban scenario

Fig. 3 Scenario creation

The scenario for urban terrain was designed on a Cartesian coordinate system with a scenario dimension of 3000 m × 3000 m. The altitude was set to 1500 m above sea level with a weather mobility interval of 10 s. The general channel properties were set as per Table 3. Each node (vehicle) has two interfaces at the physical layer, one for the control channel and the other for the service channel. The MACDOT11p extension integrated with EXata and the existing IP were used as the MAC and network protocols respectively. Bellman Ford routing was set as the standard routing protocol for our model. CBR was used as a typical traffic application for our work.

6 Emulation and Performance Analysis

The proposed VANET communication model was emulated in EXata with varied node densities (Vehicles). Bellman Ford routing protocol was considered as the default routing protocol. The various parameters considered for emulation are listed in Table 4.

Nodes (Vehicles) were made to move in predefined paths defined by an array of destination and arrival time of all the nodes. These are shown as flags in the scenarios. Certain nodes are kept static to consider them as RSUs. The screenshots of the emulation process captured at discrete instances of time are shown in Fig. 4.

Four separate scenarios have been created using four different node densities and emulated to evaluate the performance of the communication model. The performance metrics used for analysis are: (1) Total Unicast Data Received (bytes), (2) Unicast

Table 4 Emulation properties

Property	Value(s)
Scenario dimensions	3000 m × 3000 m urban terrain
Emulation time	500 s
Node densities	10, 20, 30, 60
Node mobility	Predefined mobility (using flags)
PHY layer	802.11 P CCH
MAC layer	802.11 P
Routing	Bellman Ford
Traffic application	CBR application
Packet-size	512 bytes

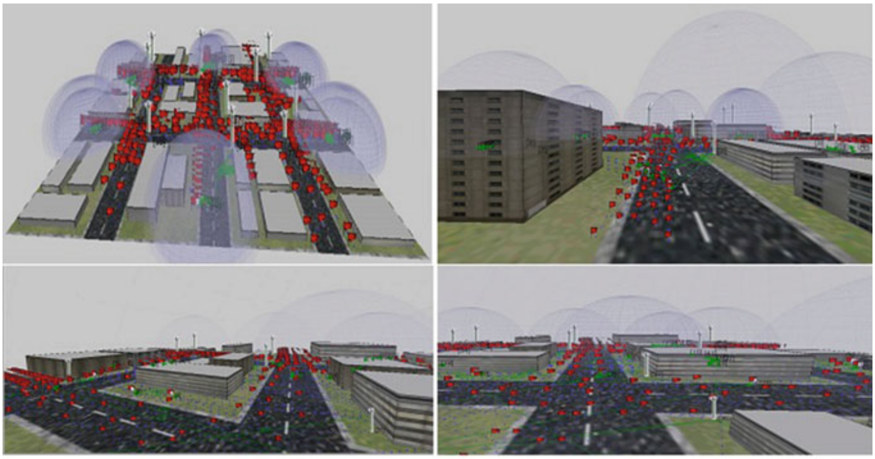


Fig. 4 Screenshots of the emulation process at different instances of time

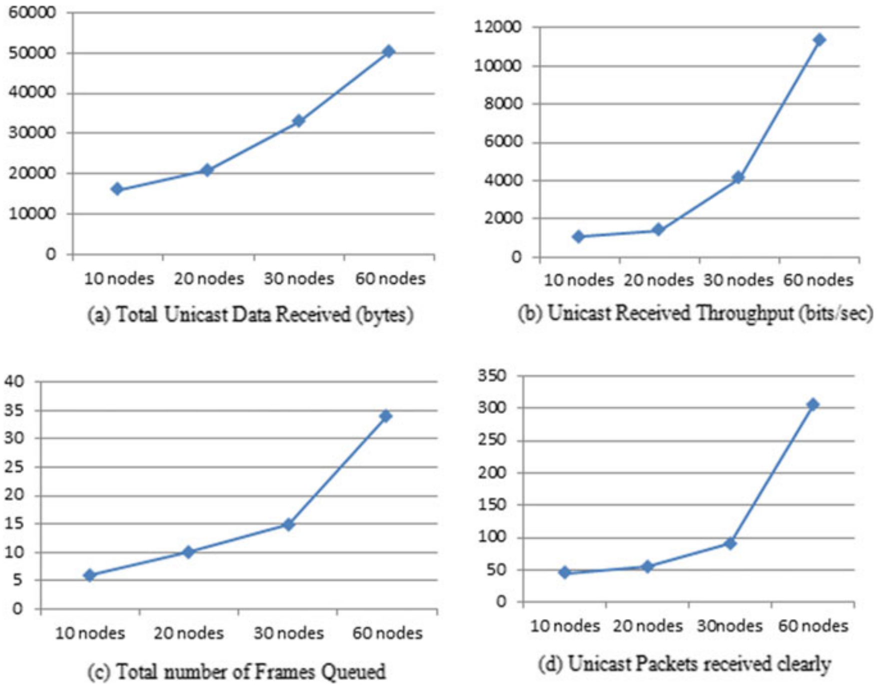
Received Throughput (bits/sec), (3) Total number of Frames Queued and (4) Unicast Packets received clearly. Our work had resulted in four trace files which were analyzed using EXata analyzer. The results obtained were tabulated as shown in Table 5 and the graphs were plotted as shown in Fig. 5 to compare the variations in the values of the performance metrics against varied node densities.

7 Conclusion

As human lives and their commuting times are very precious, VANET applications are aimed at providing quick and reliable services to commuters in respect of safety and infotainment. VANET applications cater to the needs of commuters in making

Table 5 Performance of the model at varied node densities

No. of nodes	Total unicast data received (bytes)	Unicast received throughput (bits/sec)	Total number of frames queued	Unicast packets received clearly
10	15,936	1044	6	45
20	20,992	1392	15	55
30	32,870	4140	10	90
60	50,427	11,327	34	304



(a) Total Unicast Data Received (bytes)

(b) Unicast Received Throughput (bits/sec)

(c) Total number of Frames Queued

(d) Unicast Packets received clearly

Fig. 5 Performance metrics at varied node densities

their journey safe, on-time, and joyful. On the other hand, typical VANET characteristics pose challenges for VANET application developers. There is always a demand for scalable communication mechanisms suitable for low- and high-speed vehicles, information dissemination techniques that suit sparse and dense traffic conditions. The reliability of any VANET application strongly depends upon the robustness of the underlying communication architecture. Thus, VANET communication models shall strictly adhere to the wireless access standards and specifications laid down for vehicular communications. In this work, we have designed a basic VANET communication model adhering to IEEE 802.11p standards. The Physical and Link Layer extensions of WAVE standard have been integrated with EXata5.1 to support 802.11p communication. Necessary GUI and library level extensions were also being made to create VANET scenarios. To demonstrate and evaluate our model, urban scenarios that facilitate V2V and V2I communication with different node densities were created in EXata. The communication process was emulated with predefined mobility of nodes, CBR as the traffic application and Bellman Ford as the default routing protocol. The performance metrics were recorded and analyzed using EXata analyzer. It is found that, the unicast data received had gone up considerably at higher node densities showing that the data dissemination is happening in proportion to the number of participating vehicles. Out of this unicast data received, the number of packets received clearly (without loss of information) had increased with the increase in node density. It is also evident that the throughput of the system was low at sparse traffic and high at dense traffic. As we are concentrating on urban scenarios, the density of vehicles is almost high. On the other side, the total number of frames queued was found high in dense traffic resulting in communication delay. This need not be treated as a set back because, VANET applications are expected to provide good throughput with permissible level of delay. Thus, our basic VANET communication model facilitates both V2V and V2I communication configurations and could effectively disseminate information in urban scenarios with a soft delay constraint. Our model supports a single control channel and a single service channel for communication. In future, we will evaluate the performance of several routing protocols over the proposed communication model.

References

1. Raya, M., Hubaux, J.: The security of vehicular Ad Hoc networks. In: 3rd Workshop on Security of Ad Hoc and Sensor Networks (SASN 2005), pp. 1–11. Alexandria (2005)
2. Harsch, C., Festag, A., Papadimitratos, P.: Secure position-based routing for VANETs. In: 66th IEEE Vehicular Technology Conference, pp. 26–30 (2007)
3. Samara, G., Al-Salihy, W.A., Sures, R.: Security analysis of vehicular Ad Hoc networks (VANET). In: Second International Conference on Network Applications, Protocols and Services, pp. 55–60 (2010)
4. Stampoulis, A., Chai, Z.: A Survey of Security in Vehicular Networks. Project CPSC 534 (2007)
5. Premasudha, B.G., Ram, V.R., Miller, J., Suma, R.: A review of security threats, solutions and trust management in VANETs. Int. J. Next-Gener. Comput. **7**, 38–57 (2016)
6. Zeadally, S., et al.: Vehicular Ad Hoc networks (VANETS): status, results, and challenges. In: Telecommunication Systems. Springer, Berlin (2010)
7. IEEE P802.11p/D3.0. Draft Amendment for Wireless Access in Vehicular Environments (WAVE) (2007)
8. IEEE Standard 1609.1. IEEE Trial-Use Standard for Wireless Access in Vehicular Environments (WAVE)—Resource Manager, pp. 1–63 (2006)
9. IEEE Standard 1609.2. IEEE Trial-Use Standard for Wireless Access in Vehicular Environments—Security Services for Applications and Management Messages, pp. 1–105 (2006)
10. IEEE Standard 1609.3. IEEE Trial-Use Standard for Wireless Access in Vehicular Environments (WAVE)—Networking Services, pp. 1–87 (2007)
11. IEEE Standard 1609.4. IEEE Trial-Use Standard for Wireless Access in Vehicular Environments (WAVE)—Multi-channel Operation, pp. 1–74 (2006)

12. http://web.scalable-networks.com/exata-network-emulator-software
13. Blum, J.J., Neiswender, A., Eskandarian, A.: Denial of service attacks on inter-vehicle communication networks. In: 11th IEEE International Conference on Intelligent Transportation Systems, pp. 797–802 (2008)

Predicting the Severity of Closed Source Bug Reports Using Ensemble Methods

M. N. Pushpalatha and M. Mrunalini

Abstract Severity tells about how urgent given bug is to be fixed. There are large numbers of bug identified during software development and maintenance for each bug the bug report will be submitted. Bug report gives very important information such as Description, Severity, Priority, Date and Time of bug report, etc. Even though there are clear guidelines present about how to assign the severity, the inexperienced and busy test engineer may make the mistake in correctly identifying the severity in closed source software development. Automatic prediction of severity helps inexperienced and busy engineer in saving time and resources. In this paper, bug report dataset (PITS) is taken for NASA projects from PROMISE Repository. Predicting the severity is done using Bagging, Voting, Adaboost and Random forest ensemble methods. The result shows bagging gives better accuracy than other ensemble algorithms. Two preprocessing techniques, i.e., Information gain and Chi-square are considered for data reduction. Information gain gives slightly good accuracies over chi-square.

1 Introduction

During open source software development lots of bug reports are submitted by end-users, developers, testers from all over the world are stored in different open source bug repositories like Bugzilla, JIRA, etc. lot of research is done on these repositories to analyze the bug reports for improving the quality of software and delivering the software according to customer requirements. Some of the research is related to predicting the severity of bug reports and predicting the proper developer for fixing that bug. Closed source software developments use slightly different approach for software development than open source. In open source anyone (users, developers,

M. N. Pushpalatha (✉) · M. Mrunalini
Ramaiah Institute of Technology, Bangalore 560054, India
e-mail: pushpalathamn@msrit.edu

M. Mrunalini
e-mail: mrunalini@msrit.edu

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_62

and testers) can test the software and if they find any defect then can submit the bug report to the developer. Even though there are clear guidelines on how to assign the severity. Many a times the users make the mistake when assigning the severity to the bug report. In closed source, the assignment of severity is done by test engineer who tests the software. If the test engineer is busy or inexperienced then they are the chances of making the mistakes. Prediction of severity of bug report for closed source will help for the inexperienced and busy test engineer. It is very important to identify it correctly for resource allocation and fixing urgent and critical bug. Assessment of severity in closed source depends on the experience of test engineer and the time he spends on the defect report [1].

In [1–6] used the general classifier such as rule based classification, Naïve Bayes, Naïve Bayes Multinomial, K-Nearest Neighbor, J48, RIPPER, probability based Naïve Bayes, Random Forests and Support vector machine, etc., for predicting the severity. Bagging ensemble method used in [7] for predicting the severity for open source data sets and compared with C4.5. Bagging gave good accuracy over C4.5 general classifier. Literature shows that ensemble methods are not addressed in available work for open source NASA datasets.

In this paper, the defect reports of the NASA's Project Issue and Tracking system (PITS) from PROMISE repository are considered for experiment. PITS is database contains all findings which are captured during NASA's Independent Verification and Validation (IV & V) and contains data for more than 10 years [1]. It contains the data about nuclear reactor, robotics and human-rated missions.

In this paper, different ensemble methods used for predicting the severity for PITS defect reports.

The paper organization is Sect. 2 explains the literature survey, Sect. 3 is about methodology used, in Sect. 4 Result and Discussion is presented and Sect. 5 is about the conclusion and future work.

2 Literature Survey

In the literature machine learning and Text mining techniques are used to address the different problems on bug tracking repositories. Some of the problems addressed by different researchers are on predicting the duplicate bug reports, predicting the severity and priority of bug reports and predicting the developer to resolve the bug, etc. In author [8] used natural language processing to detect duplicate defect reports. In the presence of ancillary data about a bug (e.g., number of affected users), the process of bug triaging could be automated. In this vein, Naive Bayes based classification algorithm has been used to automatically predict the severity of reported bugs [2] of Bugzilla repositories for Eclipse, Mozilla and GNOME component.

Since bug reports typically come with textual descriptions, text mining techniques have been applied on the descriptions of bug reports to automatically triage bugs [9, 10].

The predicting the severity levels for closed source of NASA defect reports is done using RIPPER algorithm [3]. Different measures like recall, precision, and F-measure is used for evaluating the result.

Prediction of severity of open source bug reports from Bugzilla is done by using Naïve Bayes Multinomial, K-Nearest Neighbor, Naïve Bayes, and Support vector machine [3]. Among four algorithms [3] found Naïve Bayes Multinomial gives good accuracy and works with less training sets and fastest. Nearest Neighbor algorithm is used in [4] for predicting the severity of open source software bug reports of Eclipse, OpenOffice and Mozilla from Bugzilla repository. In [5] author used Naïve Bayes Multinomial, Support Vector Machine, Naïve Bayes, k-Nearest Neighbor, J48 and RIPPER algorithms are used for predicting the severity of NASA defect reports, accuracy and F-measure is used for evaluating the result. In author [6] taken NASA's defect reports from PROMISE repository as Closed source data set and bug reports of Eclipse, Mozilla & GNOME from Bugzilla bug repository as open source data sets and used different classification algorithm such as Random Forests, RIPPER, Naïve Bayes, Support Vector Machine and J48 for predicting the severity of both open source and closed source datasets.

Cross projects severity prediction of bug report is done using K-NN, Naïve Bayes and Support vector machine. K-NN gave good performance over other two [11]. For dealing with imbalance bug data problem used the vote and bagging ensemble methods from RapidMiner. F-measure Performance was increased by 5% and 10% using vote and bagging respectively [11]. In this paper used the voting, bagging, Adaboost and random forest ensemble methods from RapidMiner for predicting the severity of closed source data sets.

In [12] Bayesian Networks, Naïve Bayes, REPTree, SVM, Decision tree, rules and Random Forest machine learning algorithm along with Stacking ensemble method for predicting the developer for industrial data and comparison is done on different classification algorithm and concluded that stacking ensemble method increased the accuracy. In [13] authors used bagging ensemble method and Naïve Bayes general classifier for predicting the developer of open source projects and concluded that accuracy can be increased with bagging ensemble method. In this paper, ensemble methods are used for closed source software bug reports.

3 Methodology

NASA's PITs Datasets are taken from the Promise repository [14] and used RapidMiner tool for prediction of severity. NASA's Independent Verification and Validation facility given the five anonymous PITs projects named it as pitsA to pitsF and all five projects are related to robotic.

Table 1 shows the number of bug reports available for each severity and Table 2 tells about the total number of bug reports available for each datasets, their size and total word count is given

Table 1 Number of bug reports for each severity

Projects	Severity 1	Severity 2	Severity 3	Severity 4	Severity 5
PitsA	0	325	375	239	26
PitsB	0	23	523	382	59
PitsC	0	0	132	180	7
PitsD	0	1	167	13	1
PitsE	0	24	517	243	41
PitsF	0	9	477	209	48

Table 2 Total number of bug reports, size and word count

	Total number of reports	Size of the dataset	Total word count
PitsA	965	1.2 MB	173,963
PitsB	987	704.1 KB	104,052
PitsC	319	143.6 KB	23,799
PitsD	182	106.5 KB	15,868
PitsE	825	650.0 KB	93,750
PitsF	743	548.9 KB	82,775

PITS dataset is preprocessed before applying the classification algorithm. The dataset is first tokenized to splits the text of a document into a sequence of tokens. After stop words removal is used to remove the stop words like a, the, etc., next porter stemming is used to stem the words for example the present, presented and presenting is stemmed to present. The dimensionality is reduced to 150 by using Chi-Squared Statistic and information gain, next different ensemble methods are applied on the reduced data set for classification. From Table 2 shows that number of words for each dataset which varies from 15,868 to 1,73,964. Dimensionality is reduced in order to reduce both time and memory taken by data mining algorithms.

Bagging classifier is created using K-NN classifier and in Adaboost is created using the Naïve Bayes as base classifier. In vote used Naïve Bayes, decision tree and K-NN as base classifier and majority vote from three classifiers is considered as class.

Chi-squared

This is a preprocessing technique used for term reduction. Chi-squared is used for calculating the relevance of the terms with respect to class attribute. The term is more relevant if has higher weight. It is used only for nominal label.

The Chi-square is calculated using below equation [1].

$$X^2 = \text{Sigma}[(O - E)^2/E] \tag{1}$$

In Eq. (1) X^2 is the chi-square statistic, the observed frequency is O and the expected frequency is E. The chi-squared statistic summarizes the divergence between the expected number of times each result occurs and the observed number of times each result occurs, by summing the squares of the variation, normalized by the expected numbers, over all the categories [15].

Information gain

Information gain is another preprocessing technique used for dimensionality reduction. The information gain is used for calculating the relevance of attributes based on the weights. The term is more relevant if it has highest weight. It calculates the weight of the terms with respect to class attribute. It can only be applied to nominal label [15].

3.1 Adaboost

The most popular boosting algorithm is Adaboost. There are data sets of D of d class-labeled records, $(A_1, c_1), (A_2, c_2), \ldots (A_d, c_d)$, Where c_i is the class label of record A_i. An equal weight of $1/d$ is assigned to each training record.

k rounds are required to generate k classifiers. In round i, the records from the D are sampled to form a training set, D_i, of size d.

The same sample may be selected more than once because the sampling with replacement is used. Based on the weight of sample is selected. The p classifiers are generated in p rounds. Training set D_i of size d is formed the samples of the D in round i. The classifier model M_i is created by using training samples of D_i. Test set D_i is used for calculating the error. If a sample is classified incorrectly then weight is increased otherwise weight is decreased. For generating the training records for the next round weights will be used. More focus is given on the misclassified samples of the previous round [16].

3.2 Bagging

Bagging is also known as Bootstrap aggregating is an ensemble classification technique, which combines the voting from multiple models. Multiple models are of same type. Over fitting can be avoided and also variance can be reduced using bagging [15].

3.3 Random Forest

Random forest is constructed using multiple decision trees or random trees. Each random tree is created using a random subset of features at each split, except this remaining everything is similar to decision tree [15]. It works well if data sets contain

Table 3 Accuracy of ensemble classifier using weight by Chi-squared statistic

	Bagging	Random forest	Voting	Adaboost
PitsA	75.33	56.23	72.93	58.10
PitsB	80.84	48.97	80.48	66.54
PitsC	89.79	79.88	90.10	78.96
PitsD	96.20	92.89	95.64	92.87
PitsE	66.42	63.15	69.45	40.26
PitsF	76.10	64.64	69.45	64.10

more redundant attributes [17]. New test data us classified based on the vote it receives from the multiple random trees. Suppose, if random forest is created using 10 random trees. If 8 random trees classifiers assign class as 4 and remaining two random trees as class 5, then it will be classified as class 4 because of majority votes.

3.4 Voting

Voting ensemble method is present in RapidMiner tool [15]. This method uses a majority vote for classification from the base classifiers provided. Base classifiers can be of different types. Suppose if there are three base classifiers it, if two base classifiers assign class as 3 and another one as 2. It will classify it as severity class 3. Majority vote is 2.

4 Result and Discussion

It will take more time and memory for data mining algorithms to work on huge dimensions (words). That is reason, reduced the dimension to 150 by using two dimensionality reduction methods, i.e., Chi-Square and Information gain. Table 3 shows the accuracies of different ensemble methods after reducing the dimension using Chi-Squared statistic. For PitsA Accuracy varies between 56.23 and 75.33, PitsB accuracy varies between 48.97 and 80.84, PitsC between 78.96 and 90.10, PitsD varies from 92.87 to 96.20, for PitsE varies between 40.26 and 69.45 and PitsF accuracy varies between 64.10 and 76.10. Table 4 shows the accuracies of different ensemble methods after reducing dimensionality using Information gain. For PitsA Accuracy varies between 58.09 and 74.07, PitsB accuracy varies between 54.38 and 80.72, PitsC varies between 79.85 and 89.80, PitsD varies between 93.42 and 96.20, PitsE varies between 69.21 and 72.36 and PitsF accuracy varies between 64.10 and 75.70 using different ensemble methods.

Table 4 Accuracy of ensemble classifier using weight by information gain

	Bagging	Random forest	Voting	Adaboost
PitsA	74.07	58.09	71.27	61.01
PitsB	80.72	54.38	80.35	72.20
PitsC	89.80	79.85	89.79	84.22
PitsD	96.20	93.42	95.64	92.87
PitsE	69.21	63.76	72.36	47.39
PitsF	72.86	64.64	75.70	64.10

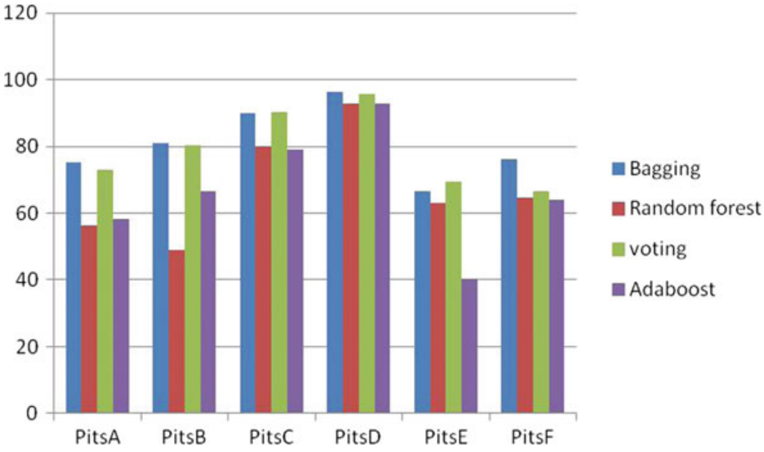


Fig. 1 Accuracies comparison using weight by Chi-squared statistic

Graphical representation of accuracies comparison is shown in Figs. 1 and 2 using Chi-Squared Statistic and Information gain respectively. Figures 1 and 2 show that bagging is given good accuracies over other ensemble methods.

Figure 3 shows accuracies comparison each classifier with different dimensionality reduction, information gain gives slightly good accuracies comparing to Chi-Squared Statistic. Accuracies of bagging and voting algorithm is same, only slight differences in the accuracies of adaboost and Random forest after reducing the dimension using Information gain and Chi-Square.

5 Conclusion

In this paper, predicting the severity of bug report for closed source dataset done using the different ensemble methods such as Bagging, Voting, Adaboost, and random forest. In that bagging is given the good accuracy over oth**er** methods. Also compared the two techniques of dimensionality reduction, i.e., chi-square and information

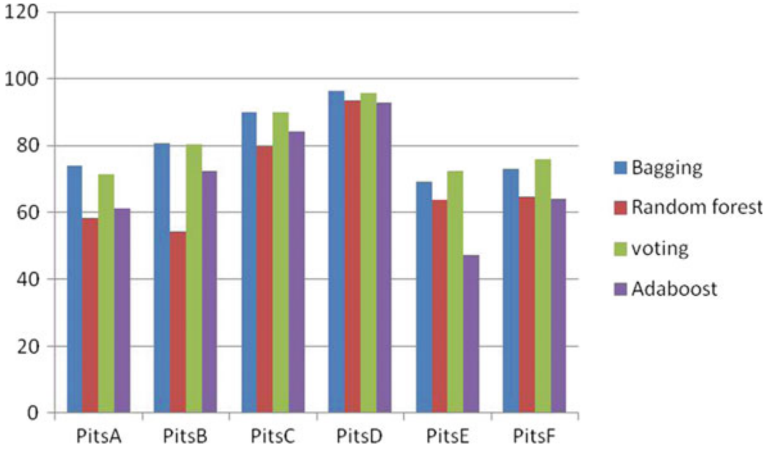


Fig. 2 Accuracies comparison using weight by information gain

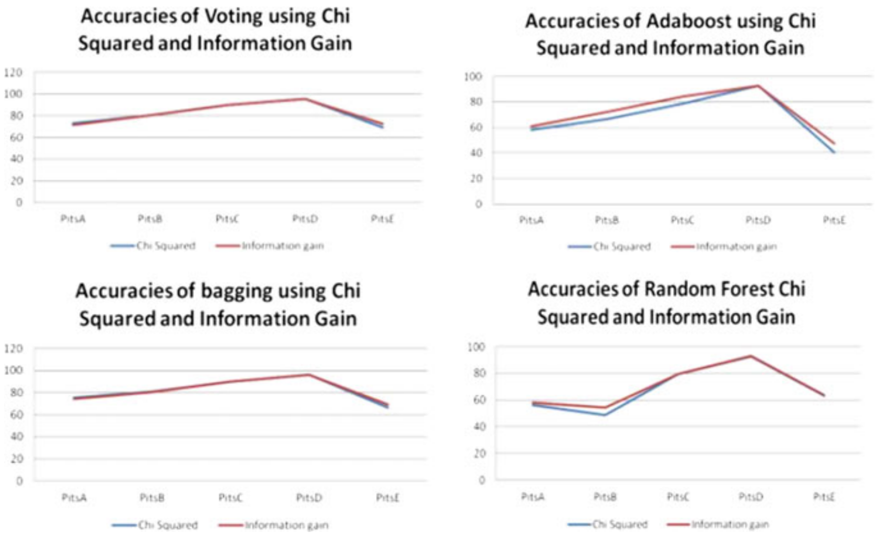


Fig. 3 Accuracies of different ensemble methods using Chi-squared and information gain

gain for reducing the number of dimension. Information gain is given slightly good accuracy over the chi-square. Better prediction of severity for NASA defect reports can be done using ensemble methods which help for improving the quality of software and on time delivery. Future work is done on the data sets of open source software for cross project context.

References

1. Menzies, T., Marcus, A.: Automated severity assessment of software defect reports. In: IEEE International Conference on Software Maintenance, vol. 28, 4 October 2008, pp. 346–355
2. Lamkanfi, A., Demeyer, S., Giger, E., Goethals, B.: Predicting the severity of a reported bug. In: 7th IEEE Working Conference on Mining Software Repositories (MSR 2010), pp. 1–10 (2010)
3. Lamkanfi, A., Demeyer, S., Soetens, Q.D., Verdonck, T.: Comparing mining algorithms for predicting the severity of a reported bug. In: 2011 15th European Conference on Software Maintenance and Reengineering
4. Tian, Y., Lo, D., Sun, C.: Information retrieval based nearest neighbor classification for fine-grained bug severity prediction. In: 2012 19th Working Conference on Reverse Engineering
5. Chaturvedi, K.K., Singh, V.B.: Determining bug severity using machine learning techniques. In: 2012 CSI Sixth International Conference on Software Engineering (CONSEG), pp. 1–6 (2012)
6. Chaturvedi, K.K., Singh, V.B.: An empirical comparison of machine learning techniques in predicting the bug severity of open and closed source projects. In: International Journal on Open Source Software and Processes, vol. 4, issue 2, pp. 32–59 (April 2012)
7. Pushpalatha, M.N., Mrunalini, M.: Predicting the severity of bug reports using classification algorithms. In: International Conference on Circuits, Controls, Communications and Computing (I4C), pp. 1–4 (October 2016)
8. Runeson, P., Alexandersson, M., Nyholm, O.: Detection of duplicate defect reports using natural language processing. In: Proceedings of the 29th International Conference on Software Engineering (ICSE ‘07), pp. 499–510, May 20–26 (2007)
9. Cubranic, D., Murphy, G.C.: Automatic bug triage using text categorization. In: Proceedings of the Sixteenth International Conference on Software Engineering & Knowledge Engineering, pp. 92–97 (June 2004)
10. Anvik, J., Hiew, L., Murphy, G.C.: Who should fix this bug? In: Proceedings of the 28th International Conference on Software Engineering (ICSE ‘06), pp. 361–370. ACM, New York (2006)
11. Singh, V.B., Misra, S., Sharma, M.: Bug severity assessment in cross project context and identifying training candidates. J. Inf. Knowl. Manage. **16**(01) (March 2017)
12. Jonsson, L., Borg, M., Broman, D., Sandahl, K., Eldh, S., Runeson, P.: Automated bug assignment: ensemble-based machine learning in large scale industrial contexts. Empirical Softw. Eng. **21**(4), 1533–1578 (2016)
13. Pushpalatha, M.N., Mrunalini, M.: Automatic bug assignment using bagging ensemble method. Int. J. Adv. Inf. Sci. Technol. **40**(40) (August 2015)
14. Promise. http://openscience.us/repo/issues/
15. RapidMiner. https://rapidminer.com/
16. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaugmann (2006)
17. Tan, P.-N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Pearson Education (2006)

Survey on Sentiment Analysis from Affective Multimodal Content

Sujay Angadi and R. Venkata Siva Reddy

Abstract For sentiment analysis, online generated user content is important particularly in social media. People are using online media increasingly to say their opinions and share their knowledge's through videos, tweets, and audio recordings. Analysis of such extensive content can help to extract user sentiments towards events or topic. With the advancement of social network for sharing reviews, recommendations, feedback, opinions, and ratings it has become a necessary for analysis. The emotions and sentiments helps in making critical decisions in organizations and businesses, and also situation awareness in environment for individuals. Since shared content on social media is multimodal in nature, research in affective computing has evolved over development of multimodal analysis frameworks. Natural language processing for text and emotion recognition from audio and visual is practice for analysis. This paper presents the overview of different techniques and approaches in sentiment analysis for text, audio, and visual modalities.

1 Introduction

Sentiment analysis has become recent progression in social media analytics and it is helping to know the emotions expressed on different stage. With the enhancement of the technology, fast growing of social media, large amount of data is being added as video, text and as audio clip. For example, customers can record their opinions and reviews on particular products by using the camera in phones and upload instantly on social media like Facebook, YouTube, etc., to report subscriber about their views.

S. Angadi (✉)
School of Computing & Information Technology, Reva University, Bangalore,
Karnataka, India
e-mail: sujayangadi90@gmail.com

R. Venkata Siva Reddy
School of Electronics & Communication Engineering, Reva University, Bangalore,
Karnataka, India
e-mail: rvsiva_reddy@ieee.org

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_63

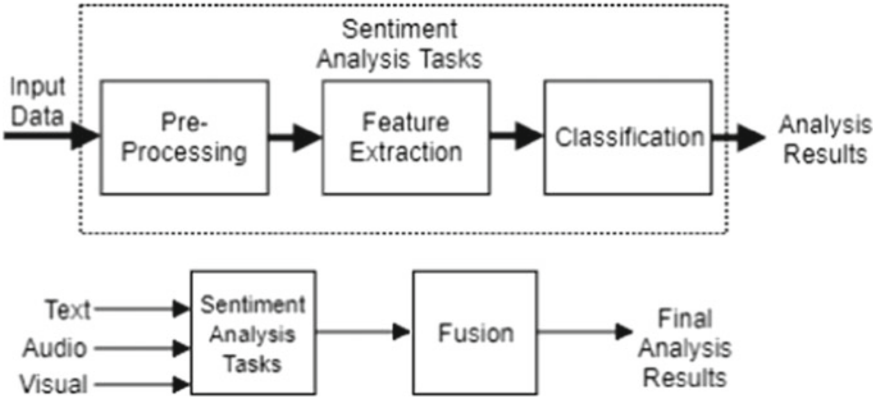


Fig. 1 Sentiment evaluation from affective multimodal content

These opinions from the customers can help organizations to improve their product or business.

With improving accessibility and regards opinion rich resources like social networking sites and private blogs, new chance and challenges arises. Usage of information technologies to find out and know the opinion of people has become a challenge in the field of affective computing.

People look beyond just rating of an product or any service. Consideration of only the overall ratings, it fails to adequately represent the other multiple potential dimensions on which any entity can be reviewed [1].

With the availability of different social media platforms one expresses his/her opinion in different modality. In such conditions it becomes necessary to have a system that performs analysis on all modalities. Since majority of the research works are extensively carried out on single modality the development of multimodal frameworks has gained attention of researchers in recent past.

Figure 1 represents the sentiment evaluation from the affective multimodal content. Here the inputs are text, audio, and video datasets. Preprocessing of input in terms of enhancement, restoration or just proper representation of data is a first step. Next step is to extract the features from all segments of audio, video, and text to take the average and compute the final feature vector. Later, separate supervised classifiers for each modality on the feature vectors are applied. Finally, overall sentiment expressed by multimodal content is identified by fusing results of analysis from textual, audio and visual data.

2 Literature Survey

In this section, we present the literature survey on the sentiment analysis from affective multimodal content, to analyze the sentiments or emotions expressed by the people. This section presents various techniques to analyze the sentiments from text, audio and visual data. Also discusses few multimodal frameworks for analysis.

2.1 Sentiment Analysis in Text

Bin et al. [1] they introduced an weakly supervised approach that uses just minimum prior information in the seed word form. For multi-aspect rating prediction author proposed usage of overall ratings in conjunction with sentence labeling. For multi-aspect sentiment analysis the role of unsupervised and weakly supervised topic representation approach is examined. They express that on multi-aspect sentence labeling task weakly supervised performs well. And can be utilized to assist multi-aspect rating calculation with only indirect supervision. Lastly, they reported that integrating characteristics obtained from unsupervised topic methods gives considerable improvement in presentation.

Singh et al. [2] introduced experimental study on a domain specific feature-based heuristic for the aspect-level sentiment analysis of movie reviews. They have developed an aspect based scheme that examines the textual reviews of the movie and given a sentiment label on every aspect. They have utilized the SentiWordNet (SWN) based system along with other selections of linguistic feature including of verb, adverb and adjective and n-gram feature withdrawal. They evaluated the result with the result gained using the Alchemy API. The result gained express that their system creates better sentiment profile than the normal document-level sentiment analysis.

María del Pilar et al. [3] proposed aspect-level sentiment analysis on tweets related to diabetes. N-gram methods such as N-gram after, N-gram before and N-gram around are used to identify the closest words to diabetes in tweets. Finally the polarity of these identified words is calculated by using SWN. It is reported by author that N-gram around has higher results than N-gram before and N-gram after. Thus author concludes that "N-gram around" method represents an optimal means to conduct sentiment analysis on tweets related to diabetes and particularly in English language.

2.2 Emotion Recognition in Audio

Zou et al. [4] proposed method for emotion recognition using Deep Belief Nets (DBN). It was projected for use of emotion information hidden in speech spectrum diagram—Spectrogram as image features. To increase subset of feature subset and to enhance detection, along with traditional features, the spectrogram features were extracted from color, orientation and brightness. Alternatives of DBN are also proposed along with approximating optimum feature subset. The proposed system was experimented on ABC (Airplane behaviors Corpus) database and Chinese corpora.

Nancy et al. [5] focuses on verifying the emotional state from the speech signal. Different audio characteristics are taken form short term and overlapping frames obtained from the speech signal. To choose a subset of helpful characteristics from full candidate feature vector, sequential backward selection (SBS) method is utilized with K-fold cross confirmation. Recognition of sentiments in the sample is completed by either Linear Discriminate Analysis (LDA) classifier or a pre-trained Support Vector Machine (SVM) model. This approach is experienced with two operated emotional database—RML Emotional Database (RED) and Berlin Database of Emotional Speech (EmoDB).

Mirsamadi et al. [6] studies the utilization of deep learning to robotically determine emotionally applicable features from the speech. It is expressed that opportunity for learning short-time frame level audio features related to emotions and also suitable temporal aggregation of features into a compact representation in utterance level using deep Recurring Neural Networks (RNN). Furthermore, they introduce a novel plan for pooling feature over time which utilizes local concentration in order to concentrate on definite sections of a speech signal which are much sensitively salient.

Tashev et al. [7] examines mixtures of the GMM (Gaussian Mixture Model)-based low level extraction of feature with neural networks. Architecture benefits in adding rapid improving neural network-based clarifications with the statistical classic approaches affected to detection of the emotions. They have evaluated the presentation of some GMM-based algorithms, which approximate the values of the whole speech first and second they complete classification. With their state-of-the-art DNN-ELMK algorithm (deep neural network- Kernel Extreme learning machine) outperforms other GMM-based algorithms.

2.3 Emotion Recognition in Visual Data

Lopes et al. [8] proposed a normal solution for the recognitions of the facial expression that utilizes the mixture of standard techniques, like specific image preprocessing and convolutional network steps. It is reported that specific combination of normalization techniques can improve recognition rate. The introduced technique reaches the better result in literature, accuracy of 97.81%, and it will take lesser time to teach than the state-of-the-art techniques.

Mayya et al. [9] proposed a new technique for robotically identifying facial expressions utilizing DCNN (deep convolutional neural network) characteristics is introduced. Due to the practice of (GPGPU) general purpose graphic processing, the feature extraction time is considerably decreased. From the assessment on two widely obtainable datasets of facial expression, they have established that utilizing features of DCNN; they can reach the recognition rate of state of the art.

Yanpeng et al. [10] proposed that framework for fusing features from different ares of face. Features were extracted using local binary patterns (LBP) and histogram oriented gradient (HOG). Later feature reduction by principle component analysis (PCA). Normalizing face areas is done to make different subjects to have same size. To obtain better recognition rate gamma correction is applied on LBP features.

Multiple classifiers based on SVM like SVM-RBF, SVM-Polynomial and SVM-Linear are studied. Proposed system is experimented on CK+ and JAFFE database.

2.4 Multimodal Sentiment Analysis

Martin et al. [11] focused on analysis of videos where reviews of different movies are expressed. For spoken utterance, Bidirectional Long Short-Term Memory (BLSTM) gives score and SVM predicts final sentiment. Authors have considered a novel database named Multi-Modal Movie Opinion (ICT-MMMO) and Metacritic corpus for evaluation. Comparative analysis is performed on usage of manual transcription and Automatic Speech Recognition (ASR) for linguistic analysis. Authors have performed cross domain, in-domain and open domain analysis. They have reported that audio-visual analysis results are still effective than when textual information is included.

Verónica et al. [12] introduces a technique for classification of multimodal sentiment, which can find out the sentiment expressed in utterance level. They presented a multimodal opinion Utterances dataset (MOUD) containing of sentiment explained utterances taken from reviews of video, where every utterance is connected with an acoustic, video, and linguistic data stream. The result of the research shows that categorization of the sentiment can be efficiently presented on multimodal data stream.

Verónica et al. [13] shows a technique that combines audio, visual and linguistic features for the reason of finding sentiment in online videos. Experimentation was concentrated on a dataset containing of Spanish videos taken from social media website YouTube. Although relative experiments, they express that the combined use of audio, textual and visual features significantly progresses over the use of only one modality at a time. Furthermore, they as well as test the probability of their multimodal technique, and run estimations on second dataset of the English videos.

Poria et al. [14] introduced a novel methodology for collecting sentiments from the videos which represents the audio, visual, and textual modalities as a source of the information. They have used two fusion methods, which are decision and feature level fusion methods to combine sentimental information taken from the multiple modalities. A detailed assessment with active works in this area is carried out through the paper, which express the novelty of their approach. Experimental results are evaluated from multiple classifiers such SVM, Extreme learning Machine (ELM) and Artificial Neural Networks (ANN). Initial relative experiments with the dataset of YouTube express that the introduced multimodal method attains nearly 80% accuracy.

Poria et al. [15] proposed convolution neural network for sentiment prediction for visual and textual data. Audio emotion recognition using features extracted through OpenSmile tool. By nourishing all features to multiple kernel learning (MKL) classifiers, they considerably outperform the state of the art of multimodal emotion detection and sentiment analysis on variety of datasets.

Table 1 illustrates comparative study of works listed in our literature survey.

Table 1 Comparative study

References	Year	Modality	Algorithms used	Dataset	Findings
[1]	2011	Text	Weakly supervised approach	Reviews from Opentable.com, Citysearch.com & trip advisor	Better performance over fully supervised baseline
[2]	2013	Text	SWN based system and classification: document-level & aspect-level sentiment classifier	Movies reviews from imdb.com	SWN (AAAVC) performs with good accuracy than the SWN (AAC) and Alchemy API
[3]	2017	Text	N-grams for aspect identification & polarity detection using SWN	Twitter dataset of diabetes	N-gram around performs better than N-gram before and N-gram after
[4]	2016	Audio	Feature level fusion using deep belief networks	Mandarin database	Recognition result on cross-corpus distinctly advances by 8.8%
[5]	2017	Audio	SBS for feature selection & SVM and LDA for classification	EmoDB & RED	SVM multiclass performed well than LDA
[6]	2017	Audio	Deep RNN	IEMOCAP corpus	Better performance over SVM based recognition
[7]	2017	Audio	GMM and neural network-based algorithms	Mandarin utterances from Microsoft spoken language system	DNN-ELMK performs better than GMM-based algorithms
[8]	2015	Visual	CNN & Gaussian distribution	CK+	Accuracy of 97.81%
[9]	2016	Visual	DCNN for face feature extraction & SVM for classification	CK+ & JAFFE datasets for evaluation	Recognition rate of 98.12% on JAFFE and 97% on CK+
[10]	2017	Visual	LBP + HOG	Cohn-Kanade (CK+) & JAFFE	Recognition rate of 98.3% on CK+ & 90% on JAFFE

(continued)

Table 1 (continued)

References	Year	Modality	Algorithms used	Dataset	Findings
[11]	2013	Audio-visual-text	SVM & BLSTM for final sentiment prediction	ICT-MMMO	Language independent audio-visual is still effective even if text is not considered
[12]	2013	Audio-visual-text	SVM for final classification	MOUD dataset	Reduction of error rate up to 10.5% compared to unimodal
[13]	2013	Audio-visual-text	SVM for classification	English video and Spanish video dataset	Comparative experiments shown multimodal outperforms unimodal
[14]	2015	Audio-visual-text	Feature level fusion & decision level fusion Classification: SVM, ELM & ANN	YouTube [16]	Feature level fusion performs better and ELM classifier performs best
[15]	2016	Audio-visual-text	Convolutional multiple kernel learning (MKL) method	MOUD, YouTube & ICT-MMMO	Proposed system performed better than the state of the art

Combination of formal, informal, and personal communication in social media is still an open challenge for analysis in text along with handling relational environment. Detection of spontaneous expressions, handling partial occlusion of face and usage of other clues like head and hand movement in emotion recognition is direction towards further research in visual emotion recognition. Effective usage of audio in emotion recognition in detecting fear and sadness is greatly expressed [17], other emotional states can be explored in absence of visual. Although multimodal outperforms unimodal systems achieving accuracy, precision and recall is still a challenge. The usage of spoken text along with audio-visual and fusion still brings up issue: how to account the correlation between different modalities and components in fusion process.

3 Conclusion

Recent research on multimodal sentiment analysis is in early stage. Sentiment analysis not only occurs in reviews, but they also occur in other forms of social media such as twitter posting, blogs, forum discussion, and commentaries. So small research work has been done in this context. In this paper we presented survey on various techniques for sentiment analysis and we offered a multimodal sentiment analysis framework, which contain set of related features for visual, audio and text data, also an easy method for adding the features taken from different modalities. Textual characteristics along with Audio-Visual data play a key role to outer perform the state of the art.

References

1. Bin, L., Myle, O., Claire, C., Benjamin, K.: Multi-aspect sentiment analysis with topic models. In: Proceedings of IEEE 11th International Conference on Data Mining Workshops, pp. 81–88 (2011)
2. Singh, V.K., Piryani, R., Uddin A., Waila, P.: Sentiment analysis of movie reviews: a new feature-based heuristic for aspect-level sentiment classification. In: International Multiconference on Automation Computing, Communication, Control and Compressed Sensing (iMac4s), pp. 712–717. IEEE, New York (2013)
3. María del Pilar, S., José, M., Katty, L., Harry, L., Miguel, Á., Rafael, V.: Sentiment analysis on tweets about diabetes: an aspect-level approach. Comput. Math. Methods Med. Article ID 5140631 (2017)
4. Zou, C., Zhang, X., Zha, C., Zhao, L.: A novel DBN feature fusion model for cross-corpus speech emotion recognition. J. Electr. Comput. Eng. Article ID 7437860 (2016)
5. Nancy, S., Abhijeet, K., Sakthivel, N.: Automatic speech emotion detection system using multidomain acoustic feature selection and classification models. In: International Conference on Identity, Security and Behaviour Analysis (ISBA), pp. 1–6. IEEE, New York (2017)
6. Mirsamadi, S., Barsoum., E., Zhang, C.: Automatic speech emotion recognition using recurrent neural networks with local attention. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2227–2231. IEEE, New York (2017)

7. Tashev, I.J., Wang, Z.-Q., Godin, K.: Speech emotion recognition based on Gaussian mixture models and deep neural networks. In: Information Theory and Applications Workshop (ITA), pp. 1–4 (2017)
8. Lopes, A.T., de Aguiar, E., Oliveira-Santos, T.: A facial expression recognition system using convolutional networks. In: SIBGRAPI Conference on Graphics, Patterns and Images, pp. 273–280 (2015)
9. Mayya, V., Radhika, P.M.: Automatic facial expression recognition using DCNN. In: Procedia Computer Science, vol. 93, pp. 453–461. Elsevier, Amsterdam (2016)
10. Yanpeng, L., Yibin, L., Xin, M., Rui, S.: Facial expression recognition with fusion features extracted from salient facial areas. Sensors (4) (2017)
11. Martin, W., Felix, W., Tobias, K., Björn, S., Congkai, S., Kenji, S., Louis-Philippe, M.: YouTube movie reviews: Sentiment analysis in an audio-visual context. IEEE Intell. Syst. **28**(3), 46–53 (2013)
12. Verónica, P., Rada, M., Morency, L.: Utterance-level multimodal sentiment analysis. ACL, vol. 1 (2013)
13. Verónica, P., Rada, M., Morency, L.: Multimodal sentiment analysis of Spanish online videos. IEEE Intell. Syst. **28**(3), 38–45 (2013)
14. Poria, S., Cambria, E., Howard, N., Huang, G.B., Hussain, A.: Fusing audio, visual and textual clues for sentiment analysis from multimodal content. Neurocomputing **174**, 50–59 (2016). (Elsevier)
15. Poria, S., Iti, C., Erik, C., Hussain, A.: Convolutional MKL based multimodal emotion recognition and sentiment analysis. In: Proceedings of IEEE International Conference on Data Mining, pp. 439–448 (2016)
16. Morency, L.P., Mihalcea, R., Doshi, P.: Towards multimodal sentiment analysis: harvesting opinions from the web. In: Proceedings of the 13th International Conference on Multimodal Interfaces, pp. 169–176. ACM, New York (2011)
17. Silva, D., Miyasato, T., Nakatsu, R.: Use of multimodal information in facial expression analysis. IEICR Trans. Inf. Syst. **E81**, 397–401 (1998)

An Effective Technique for Controlling the Speed of Sensorless Induction Motor Through Modified EKF

C. Kamal Basha and S. Mohan Krishna

Abstract The major advantages of sensorless induction motor drives are reliability and low cost due to exclusion of sensors. The sensorless schemes proceed with the speed estimate and other critical parameters as a first step. In most of the recent works, extended Kalman filter (EKF) is used to estimate the speed. The major drawback in the above-mentioned method is high computational time. In this paper, a new method has been proposed for speed estimation of sensor less induction motor using wavelet based EKF. The wavelet is used to reduce the number of iterations in EKF and the computational time is reduced. The results obtained have validated the same.

1 Introduction

The induction motor construction is rugged and effective in a hostile environment and the maintenance is also comparatively less [1–3]. But, the control is complex owing to nonlinearity and coupled dynamics [4]. For speed estimation, the complex rotor dynamics has to be decoupled and reduce the hardware. Therefore, the elimination of a speed sensor saves mounting space, sensitive electronics and reduction in cost [5]. Availability and affordability of high-speed special purpose processors like DSP, allowed it to be commercialized in many industries like process control, automation, etc. [6]. Low-speed range sensorless drives not only increase the estimation bandwidth, but also are more reliable [7]. The main purpose is to ensure that it matches the dynamic performance of a sensored vector control drive [8]. The sensorless state estimation algorithms are occupying considerable research space [9–13]. These schemes generally exploit the motor model [14, 15]. The algorithms

C. Kamal Basha (✉)
Department of Electrical and Electronics Engineering, Madanapalle Institute of Technology and Science, Madanapalle, AP, India
e-mail: kamalbashac@mits.ac.in

S. Mohan Krishna
Department of Electrical and Electronics Engineering, College of Engineering and Design, Alliance University, Bangalore, Karnataka, India

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_64

involving extended Kalman filters generally exploit the concept of state estimation for nonlinear systems. The disadvantage is high computational time and space and the solution which is not optimal. Also, the filter diverges if the model is not precise.

In this paper, wavelet-based EKF is used to estimate the speed of the sensorless induction motor. The wavelet reduces the computational time of the system. The comparison with the existing model shows improved performance.

2 Literature Survey

Several other algorithms for parameter estimation exist [16–19]. Suman et al. [20] presented a novel Model reference adaptive system (MRAS) estimation scheme for sensor less control of induction motor. The drawback was the presence of torque ripples and disturbances in the steady state region. The proposed algorithm uses the reference voltage vector in which the stator flux components are used as control variables. Khan et al. [21] discussed the concept of a hybrid MRAS speed observer. The proposed algorithm minimizes the torque ripples and improves the speed performance by ensuring that the error between the estimated and the actual speed is constrained to a low value. Abbou et al. [22] made use of two artificial intelligence based algorithms for a sensorless Induction Motor drive. MRAS estimates the rotor speed. The torque, stator flux and current ripples are reduced and there is an improvement in the dynamic performance for a wide speed bandwidth. Kianinezhad et al. [23] presented a strategy for tuning the observer, in which, a separate estimation of stator resistance was introduced. The stator resistance estimate is accurate and works well even during parameter uncertainties. Messaoudi et al. [24] presents a comparison between the EKF and the MRAS. Although both have inherent noise rejection capabilities the EKF has a slightly better observation performance but, requires an accurate load torque information and occupies more computational space whereas, the MRAS strategy is flexible and easy to implement. Doan et al. [25] presented a rotor speed estimation based EKF to minimize the speed error for a wide speed bandwidth. There is lot of mismatch in critical parameters in the low speed range. To overcome it, the authors have considered simultaneous estimation of load torque and the estimated speed. Zerdali and Barut [26] discussed different fitness functions for optimized EKF's. On comparison, the most suited fitness function is obtained for motor control applications. The major problem in EKF is many iterations.

3 Proposed Methodology

By considering the above-mentioned drawback, it is suggested to control the speed of the sensor less induction motor by using EKF and wavelet transform. It is mathematically modeled as shown below:

3.1 Mathematical Model of Sensor Less Induction Motor

The mathematical structure is given below:

$$\frac{\mathrm{d}i_\mathrm{s}}{\mathrm{d}t} = \frac{1}{\sigma L_\mathrm{s}}\left(u_\mathrm{s} - R_\mathrm{s} i_\mathrm{s} - \frac{L_m}{L_\mathrm{r}}\frac{\mathrm{d}\psi_\mathrm{r}}{\mathrm{d}t}\right) \tag{1}$$

$$\frac{\mathrm{d}\psi_\mathrm{r}}{\mathrm{d}t} = \frac{L_\mathrm{m}}{L_\mathrm{r}} i_\mathrm{s} - \left(\frac{1}{L_\mathrm{r}} - j\omega_\mathrm{r}\right)\psi_\mathrm{r} \tag{2}$$

$$\sigma = L_\mathrm{s}\left(1 - \frac{L_\mathrm{m}^2}{L_\mathrm{s} \cdot L_\mathrm{r}}\right) \tag{3}$$

where, L_s is the stator inductance, i_s is the stator current, u_s the stator voltage, R_s the stator resistance, L_m is the magnetizing inductance, L_r is the rotor inductance, ψ_r is the rotor flux. The state equation used to represent the same is shown below.

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} i_\mathrm{s} \\ \psi_\mathrm{s} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\begin{bmatrix} i_\mathrm{s} \\ \psi_\mathrm{s} \end{bmatrix} + \begin{bmatrix} B_1 \\ 0 \end{bmatrix} u_\mathrm{s} \tag{4}$$

where, ψ_s is the stator flux,

$$A_{11} = \frac{1}{\sigma L_\mathrm{s}}\left(R_\mathrm{s} + \frac{R_\mathrm{r} \cdot L_\mathrm{m}}{L_\mathrm{r}}\right) I$$

$$A_{12} = \frac{R_\mathrm{r} \cdot L_\mathrm{m}}{\sigma L_\mathrm{s} L_\mathrm{r}} I - \frac{L_\mathrm{m}}{\sigma L_\mathrm{s} L_\mathrm{r}} \omega \cdot J$$

$$A_{21} = \frac{R_\mathrm{r} \cdot L_\mathrm{m}}{L_\mathrm{r}} I$$

$$A_{22} = -\frac{R_\mathrm{r}}{L_\mathrm{r}} I + \omega \cdot J$$

$$B_1 = \frac{1}{\sigma L_\mathrm{s}}$$

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, J = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$$

ω is the rotor speed of the motor.

The electromagnetic torque developed by the motor is given by,

$$T_\mathrm{e} = \frac{3}{2} P * \psi_\mathrm{s} i_\mathrm{s} \tag{5}$$

While the load torque acts as a disturbance via the mechanical relation.

$$\frac{\mathrm{d}\omega}{\mathrm{d}t} = \frac{1}{j}(T_\mathrm{e} - T_\mathrm{L}) \tag{6}$$

where, T_L is the load torque, j is the total rotor inertia.

3.2 EKF for Speed Estimation of Sensor Less IM

In the process of estimation, the EKF is the nonlinear version of the Kalman filter which produces some linearity in between the current mean and covariance. The speed estimated by EKF is controlled by producing some variations in the flux and stator current. In EKF the most important step is to predict and update the given equation. The speed of the motor was predicted and in the next step by updation of the equation the estimated speed is obtained. The following standard sets of equations are used for EKF. x_k be the state estimation, z_k which is the covariance to be estimated. w_k and v_k be the process and observation noises respectively.

$$x_k = f(x_{k-1}, u_{k-1}) + w_{k-1} \tag{7}$$

$$z_k = h(x_k) + v_k \tag{8}$$

The prediction equations used in EKF are given below.

$$\widehat{x}_{k|k-1} = f\left(\widehat{x}_{k-1|k-1}, u_{k-1}\right) \tag{9}$$

$$P_{k|k-1} = F_{k-1}P_{k-1|k-1}F_{k-1}^{\mathrm{T}} + Q_{k-1} \tag{10}$$

Update equations used in EKF are given below.

$$\tilde{y}_k = z_k - h(\widehat{x}_{k|k-1}) \tag{11}$$

$$S_k = H_k P_{k|k-1} H_k^{\mathrm{T}} + R_k \tag{12}$$

$$K_k = P_{k|k-1} H_k^{\mathrm{T}} S_k^{-1} \tag{13}$$

$$\widehat{x}_{k|k} = \widehat{x}_{k|k-1} + K_k \tilde{Y}_k \tag{14}$$

$$P_{k|k} = (I - K_k H_k) P_{k|k-1} \tag{15}$$

$$F_{k-1} = \left.\frac{\partial f}{\partial x}\right|_{\widehat{x}_{k-1|k-1}, u_{k-1}} \tag{16}$$

$$H_k = \left.\frac{\partial h}{\partial x}\right|_{\widehat{x}_{k|k-1}} \tag{17}$$

Here,

$\tilde{y}_k$	Measurement residual
S_k	Residual covariance
K_k	Kalman gain
F_{k-1} and H_k	be the jacobian matrices of the state transition and observation.

By applying the EKF explained above we obtain the estimated speed of the motor at one time. To obtain the estimated time continuously, increase the number of iteration of the above process which, in turn, would increase the computational time.

3.3 Discrete Wavelet Transform (DWT) to Estimate Speed of Sensor Less IM

The DWT, obtained on the basis of sub-band coding, is found to be yielding a fast computation of Wavelet Transform. This technique is easy for implementation. Two sets of related functions required for DWT are scaling and wavelet functions [27–29]. In this proposed method, the DWT is used for speed estimation. The mathematical model used to compute DWT for the given signal is shown in (18).

$$y[n] = \sum_{k=-\infty}^{\infty} x[k] \cdot g[n-k] \tag{18}$$

The output obtained after applying low pass and high pass filter are shown in Eqs. (19) and (20) respectively.

$$y_{\text{low}}[n] = \sum_{k=-\infty}^{\infty} x[k] \cdot g[2n-k] \tag{19}$$

$$y_{\text{high}}[n] = \sum_{k=-\infty}^{\infty} x[k] \cdot h[2n-k] \tag{20}$$

where, $x[k]$ is the original signal, $y_{\text{high}}[n]$, $y_{\text{low}}[n]$ is the output of the high pass and low pass filter, $\mathrm{g}[n]$, $h[n]$ is the half band high and half band low pass filter. The input values to the wavelet are directly given from the EKF and the speed is obtained for different time intervals

After estimation process, the next step is speed control nearer to the reference speed value. The initial step is to calculate the flux for a particular estimated speed. This is explained in Sects. 3.2 and 3.3. After calculating the flux for the particular speed, speed is computed by using (1). The poles are always constant and by adjusting the flux and current, the torque can be adjusted. The amount of flux and current needed is computed in the proposed method. By applying this flux and current to the torque equation; a new torque value is obtained which, when substituted in (6), yields the controlled speed.

4 Result and Discussions

Figure 1 shows the estimated speed of sensorless induction motor using the proposed EKF. Due to change in speed of the motor the flux also varies, that change in flux is estimated and shown in Fig. 2. Figure 3 shows the controlled speed after applying the proposed method. For that, the rated speed 1410 rpm is taken as the reference speed. The flux value obtained to control the speed of the motor is shown in Fig. 4. The estimated speed and flux obtained using the EKF method is shown in Figs. 5 and 6 respectively and both have more variations.

The computational time of our method shown in Fig. 7 is 0.005 s and for EKF it is 0.03 s. From this it is clear that it is 0.025 s lower than EKF method. This is

Fig. 1 Estimated speed using proposed method

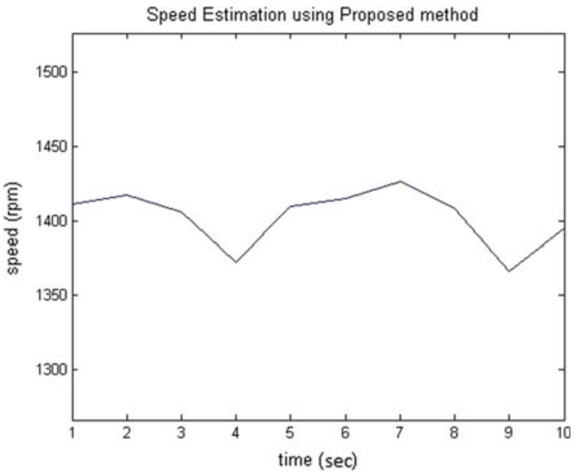


Fig. 2 Estimated flux using proposed method

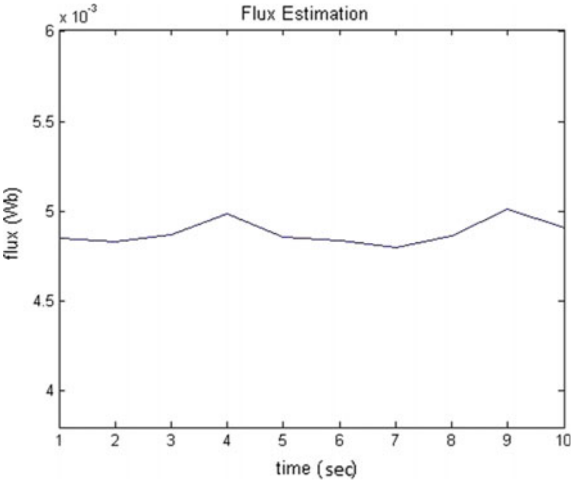


Fig. 3 Speed controlled using proposed method

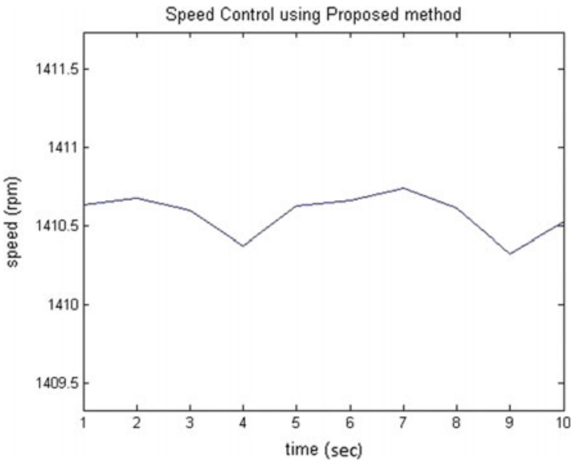
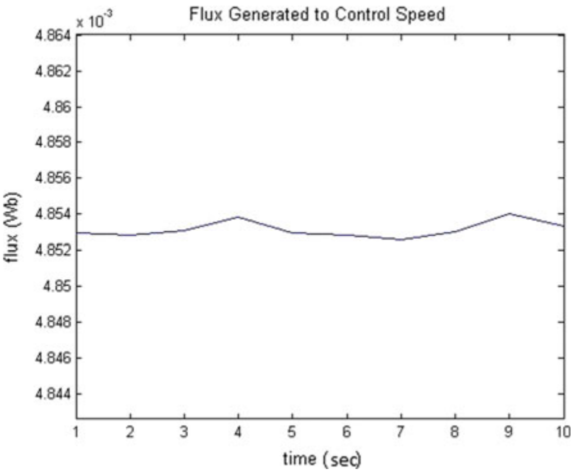


Fig. 4 Flux generated to control the speed



due to reduced number of iterations which was possible by incorporating the wavelet concept in EKF state estimation.

5 Conclusion

The speed estimated is compared with EKF method. The parameter considered for comparison of the proposed method and EKF method is computational time. The computational time obtained by using proposed method is 0.005 s and using EKF method is 0.03 s thereby, validating the improvement.

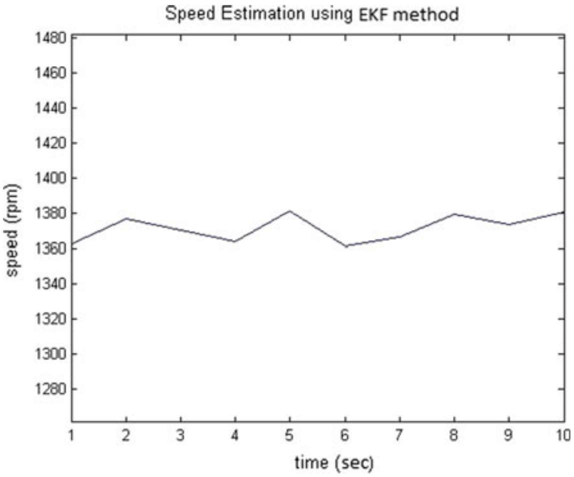


Fig. 5 Speed estimated using EKF method

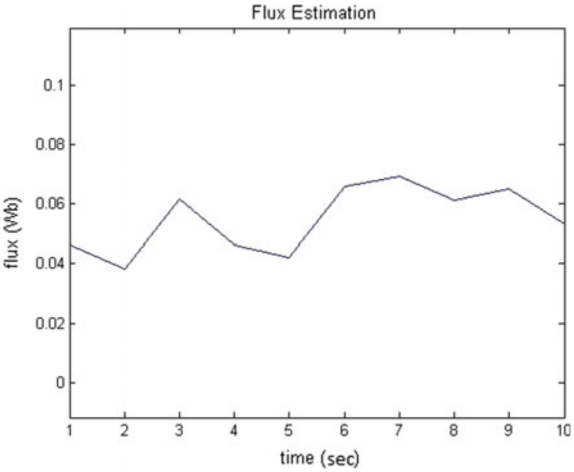


Fig. 6 Flux estimated using EKF method

Appendix

See Table 1.

Fig. 7 Computational time for both the algorithms

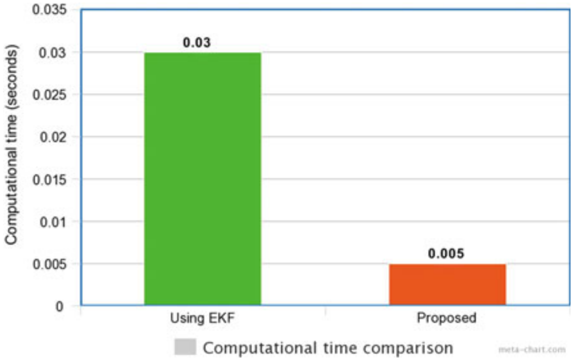


Table 1 Equivalent circuit parameters

Machine parameter	Value
R_s	19.35 Ω
R_r	8.04 Ω
L_{ls}	0.032 H
L_{lr}	0.032 H
L_m	0.601 H
J	$0.0051\ \text{kg} * \text{m}^2$
No. of pair of poles	2

References

1. Guzinski, J., Abu-Rub, H.: Sensorless induction motor drive for electric vehicle application. Int. J. Eng. Sci. Technol. **2**(10), 20–34 (2010)
2. Mishra, J.P., Gupta, S.P., Padhy, N.P.: Sensorless vector control of induction motor using direct adaptive RNN speed estimator. In: National Power Electronics Conference, pp. 1–9 (2010)
3. Barada Mohanty, K., De, N.K., Routray, A.: Sensorless control of a linearized and decoupled induction motor drive. In: National Power Systems Conference, Kharagpur, pp. 46–49 (2002)
4. Abid, M., Ramdani, Y., Meroufel, A.K.: Speed sliding mode control of sensorless induction machine. J. Electr. Eng. **57**(1), 47–51 (2006)
5. Dinesh Pai, A., Mangsuli, P.R., Rao, N.J.: Nonlinear observer based sensorless direct torque control of induction motor. In: The Third International Power Electronics and Motion Control Conference, vol. 1, pp. 440–445 (2000)
6. Mohanty, K.B.: Sensorless sliding mode control of induction motor drives. In: IEEE Region 10 Conference TENCON 2008, pp. 1–6 (2008)
7. Wolbank, T.M., Metwally, M.K.: Separation of saliency information for speed sensorless detection of induction machines flux and rotor position. Int. J. Comput. Math. Electr. Electron. Eng. **29**(5), 1380–1392 (2010)
8. El-Murr, G., Giaouris, D., Finch, J.W.: Universal PLL strategy for sensorless speed and position estimation of PMSM. In: IEEE Region 10 and the Third International Conference on Industrial and Information Systems, pp. 775–780 (2008)
9. Asseu, O., Yeo, Z., Koffi, M., Kouacou, M.A., Ali, K.E.: Sensorless control of induction machines using a reduced order extended Kalman filter for rotor time constant and flux estimation. J. Appl. Sci. **10**(5), 399–405 (2010)

10. Drevensek, D., Zarko, D., Lipo, T.A.: A study of sensorless control of induction motor at zero speed utilizing high frequency voltage injection. EPE J. **13**(3), 7–11 (2003)
11. Negadi, K., Mansouri, A., Khtemi, B.: Real time implementation of adaptive sliding mode observer based speed sensorless vector control of induction motor. Serb. J. Electr. Eng. **7**(2), 167–184 (2010)
12. Traore, D., Plestan, F., Glumineau, A., De Leon, J.: High order sliding mode control of a sensorless induction motor. In: Proceedings of the 17th World Congress the International Federation of Automatic Control, Korea, pp. 6232–6237, July 6–11 (2008)
13. El-Sawy, A.M., Mohamed, Y.S., Zaki, A.A.: Stator resistance and speed estimation for induction motor drives as influenced by saturation. Online J. Electron. Electr. Eng. **3**(2), 416–424 (2011)
14. Bensiali, E.E., Omeiri, A., Champenois, G.: Sensorless control of induction motor: design and stability analysis. World Acad. Sci. Eng. Technol. **66**, 264–269 (2010)
15. Vasic, V., Slobodan, N., Levi, E.: A stator resistance estimation scheme for speed sensorless rotor flux oriented induction motor drives. IEEE Trans. Energy Convers. **18**, 476–483 (2003)
16. Venkadesan, A., Himavathi, S., Muthuramalingam, A.: Novel SNC-NN-MRAS based speed estimator for sensor-less vector controlled IM drives. World Acad. Sci. Eng. Technol. **75**, 1212–1217 (2011)
17. Mohan Krishna, S., Febin Daya, J.L.: MRAS speed estimator with fuzzy and PI stator resistance adaptation for sensorless induction motor drives using RT-lab. Perspect. Sci. **8**, 121–126 (2016)
18. Mohan Krishna, S., Febin Daya, J.L.: An improved stator resistance adaptation mechanism in MRAS estimator for sensorless induction motor drives. Adv. Intell. Syst. Comput. **458**, 371–385 (2016)
19. Mohan Krishna, S., Febin Daya, J.L.: Dynamic performance analysis of MRAS based speed estimators for speed sensorless induction motor drives. In: IEEE International Conference on Advances in Electronics, Computers and Communication, pp. 1–6 (2014)
20. Suman, K., Aditya, V.: Sensorless control of induction motor drive using SVPWM–MRAS speed observer. J. Emerging Trends Eng. Appl. Sci. **2**(3), 509–513 (2011)
21. Haseeb Khan, Md, Amarnath, J.: Fuzzy logic based HPWM-MRAS speed observer for sensorless control of induction motor drive. ARPN J. Eng. Appl. Sci. **5**(4), 86–93 (2010)
22. Abbou, A., Mahmoudi, H.: Performance of a sensorless speed control for induction motor using DTFC strategy and intelligent techniques. J. Electr. Syst. **5**(3), 64–81 (2009)
23. Kianinezhad, R., Nahid-Mobarakeh, B., Betin, F., Capolino, G.A.: robust sensorless vector control of induction machines. Iran. J. Sci. Technol. Trans. B, Eng. **33**(B2), 133–147 (2009)
24. Messaoudi, M., Kraiem, H., Ben Hamed, M., Sbita, L., Abdelkrim, M.N.: A robust sensorless direct torque control of induction motor based on MRAS and extended Kalman filter. Leonardo J. Sci. (12) 35–56 (June 2008)
25. Doan, P.T., Bui, T.L., Kim, H.K., Byun, G.S., Kim, S.B.: Rotor speed estimation based on extended Kalman filter for sensorless vector control of induction motor. In: Zelinka, I., Duy, V., Cha, J. (eds.) AETA 2013: Recent Advances in Electrical Engineering and Related Sciences. Lecture Notes in Electrical Engineering, vol. 282. Springer, Berlin (2014)
26. Zerdali, E., Barut, M.: The comparisons of optimized extended Kalman filters for speed-sensorless control of induction motors. IEEE Trans. Industr. Electron. **64**(6), 4340–4351 (2017)
27. Meziane, S., Toufouti, R., Benalla, H.: MRAS based speed control of sensorless induction motor drives. ICGST-ACSE J. **7**(1), 43–50 (2007)
28. Sadashivappa, G., Ananda Babu, K.V.S.: Evaluation of wavelet filters for image compression. World Acad. Sci. Eng. Technol. **51**, 131–137 (2009)
29. Montanaria, M., Peresadab, S., Tillia, A., Tonielli, A.: Speed sensorless control of induction motor based on indirect field-orientation. In: IEEE Industry Applications Conference, Italy, vol. 3, pp. 1858–1865 (2000)

A Microservices-Based Smart IoT Gateway System

Pradyumna Sridhara, Narsimh Kamath and Sreeharsha Srinivas

Abstract With IoT-enabled services becoming ubiquitous, the need for its integration with cloud computing is becoming vital. IoT sensors, being of a non-standardized nature, operate over a plethora of radio frequency communication protocols and transmit data in varied packet formats, which are often incompatible with each other. In this paper, we present a microservices-based design and the details of our implementation of an IoT Gateway System that enables sensor-to-cloud connectivity across varying types of sensors, RF communication protocols, data packet formats, and even cloud service providers. The design also includes support for an application-specific edge-analytics module and a database to house sensor telemetry locally. Our proof-of-concept implementation supports 6LoWPAN for sensor communication and Azure for the cloud platform.

1 Introduction

By the year 2020, it is expected that about 50 billion devices will be connected to the internet, making it more of an Internet of Things than an Internet of People as it currently stands today, empowering cloud computing to be a Cloud of Things [1]. An IoT gateway device forms an integral part of this IoT ecosystem, acting as an intermediary between the sensor nodes and the cloud. This section broadly describes some of the key roles played by an IoT Gateway, and the challenges presented by each role.

P. Sridhara (✉)
Analog Devices India, PES University, Bengaluru, India
e-mail: pradyumnasridhar@gmail.com

N. Kamath · S. Srinivas
Analog Devices India Private Limited, Bengaluru, Karnataka, India
e-mail: Narsimh.Kamath@analog.com

S. Srinivas
e-mail: sreeharsha.srinivas@analog.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_65

1.1 Communicating with Sensor Nodes

Sensor nodes transmit telemetry using energy efficient interfaces like Bluetooth Low-Energy, ZigBee and 6LoWPAN that are based on the IEEE 802.15.4 protocol. Other common interfaces include cellular networks and Wi-Fi. The IoT gateway is required to connect to several sensor nodes and receive the data transmitted by them. However, as the sensor nodes use a variety of communication protocols, the gateway needs to abstract these in order to support multi-protocol data communication at the edge level. In addition, it should be easy to incorporate support for newer, upcoming protocols such as Wi-Fi HaLow [2] in the existing gateway system. Although there exist standards such as oBIX that have been used to provide compatibility across protocols [3], they are specific to a domain.

1.2 Parsing Sensor Telemetry

Data sent by each of the sensor nodes needs to be parsed by the gateway. The data will usually include protocol headers, metadata regarding the sensor type and most importantly the payload (sensor readings). Numerous sensor nodes can transmit temperature, light and accelerometer sensor values, each in varying packet formats. Since the interoperability standards are currently not widely incorporated, or unavailable in certain cases, it is up to the gateway to ensure its compatibility with the packet format defined by the application's specifications. Additionally, the gateway should feature the defining of parsing rules for a new sensor type's packet structure, without compromising the compatibility with the existing packet structures.

1.3 Fog Computing

Fog computing [4] refers to the extension of cloud computing services to edge devices such as gateways, to bring the power of cloud closer to where the data is being generated and processed. One of its applications is data trimming [5], where the data gathered at the sensors is filtered at the gateway before being sent to the cloud, to avoid unnecessary network traffic and usage of cloud storage space. As these services can also be application-specific, the gateway must support a customizable module that can run the application's edge-analytics logic, before the data is forwarded to the cloud. This ensures that the gateway is able to avail the benefits of the fog computing architecture while at the same time not being limited to a single use case.

1.4 Communicating with the Cloud

Cloud computing technology offers an array of services to support the IoT ecosystem, such as reliability, connectivity, computing power, analytics, and monitoring. Cloud services also allow the IoT system to scale efficiently. Numerous cloud service providers offer a diverse set of IoT solutions. This assortment of cloud services makes certain cloud platforms more suitable than others for the given application. Thus, an application-agnostic gateway must be compatible with multiple cloud service providers.

Contribution Our implementation of the IoT gateway design is able to support multiple protocols in the sub-1 GHz band to communicate with the sensor nodes. Integrating a module with the existing infrastructure to support a new protocol, involves adding just one line of code. Sensor data packets can be parsed using user-specified rules, for each type of sensor. The user can contribute application logic for edge-analytics, which the gateway will run over the sensor telemetry before the data is forwarded to the cloud. The application logic can also perform analytics over previously ingested data through a handler to the on-board database. Implementations of our design are able to successfully support data transmission across multiple cloud platforms such as Analog Devices' ADConnect and Microsoft Azure.

2 Architecture

The model consists of several independently scalable modules, each assigned to a specific task. Each module can have one or more implementations seamlessly interacting with implementations of all other modules. This section describes the structure and the flow of control and data across these modules, as depicted in Fig. 1.

Controller This central module enables all the other modules to interact and coordinate with each other. The control flow originates here. The Controller uses the services of each of the components to carry out the gateway's functionalities.

RF Driver Being responsible for receiving the data from the sensors, this module is tasked with setting up the IoT network, initializing protocol configuration parameters, initiating the handshake process with sensor nodes and most importantly, packing and unpacking messages between the gateway and the sensor nodes as per the protocol specifications. After the network is setup, the RF Driver should run perpetually, listening for incoming messages and forwarding them to the controller. This operation is asynchronous. The Controller module has an abstracted view of the RF Driver's operations. It initializes the RF Driver and waits for it to relay the data packets, which contain the payload as well as some metadata detailing the sender's specifics.

Frame Parser The data received at the Controller is forwarded to this module for parsing. The Controller invokes the services of the Frame Parser for each packet

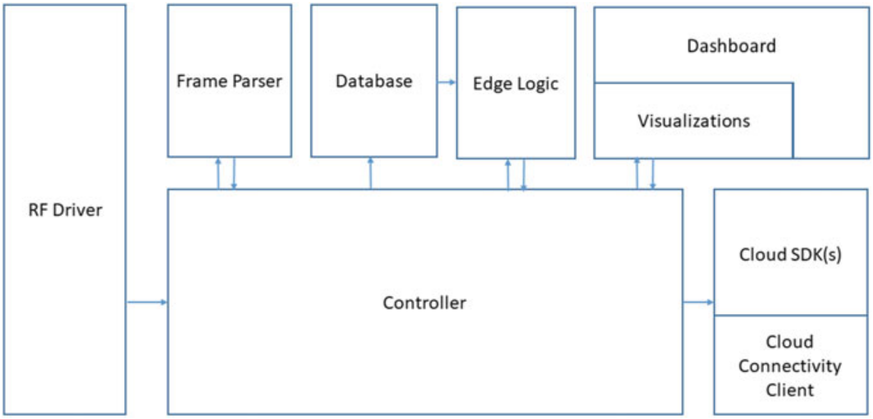


Fig. 1 The microservices-based architecture depicted along with the data flow between the components

it receives, as it is an on-demand process unlike the RF Driver. The Frame Parser consults its parsing rules table and the metadata in the packet to know how the payload is structured. This parsed data is then sent back to the Controller as key-value pairs.

Edge-Logic This module has two important roles, which in synergy provide fog computing services—a continually running background process that performs analytics over past data, and the other, that acts on a single data unit, i.e., each data packet. The latter can be used to perform feature reduction based on sensor values from multiple sensor types in real time. A handler to the database is provided to access previous data and easily execute queries. The Controller feeds each parsed data packet to this module, and receives the dependent parameters along with any other additional data that the edge-logic has computed. The implementation of this module's design is on a per-application basis.

Database The Controller stores the parsed data in a database so that the edge-logic module has access to past sensor data. Metadata pertains to the type of sensors present in each sensor node and their IDs.

Cloud Connectivity Client and Cloud SDK The Controller passes on the sensor telemetry data, along with the additional data computed on-board by the fog computing module to the cloud. A publish/subscribe design principle is used to communicate with the cloud. SDKs developed by the cloud service provider are used to register the devices on the cloud, manage service instances and generate cloud endpoints. These work in conjunction with the cloud connectivity client module.

Dashboard and Visualizations A HTTP server hosted on the gateway is used to serve the user interface of the dashboard. Various configuration settings pertaining to each module are controlled from here. All connected devices are monitored, and the sensor data is graphically represented in real time using visualization applications.

Additionally, status and error messages and logs can be seen on the dashboard. A visual representation of the database can be also be viewed. Implementation of the edge-analytics logic can be uploaded through the dashboard.

3 Module Interface Design and Abstraction

Abstracting the implementation through a common interface allows each module to scale independently, making it compatible across implementation specifications and easily adaptive to newer technologies. This approach also provides the option of disabling certain services for some use cases, such as not using the cloud services module while testing a newly developed sensor and employing the dashboard visualizations instead. This section describes our APIs and delineates its implementation particulars. Each module has its own configuration parameters which are read when the modules are initialized (through the initialize() function call).

3.1 Frame Parser

Our data packet format identifies each packet as either a sensor data packet, or a sensor registration packet. The first bit of the packet is used to make this classification. Each sensor is uniquely identified in a sensor node by its sensor ID. Similarly, each sensor type is associated with an ID. Every sensor data packet has an RTC time stamp field.

A sensor registration packet is used to register a new sensor with the gateway device, as belonging to a sensor node. It contains the sensor type, the number of data types (Eg: an accelerometer sensor has three data types—a floating point number for each of the axes), followed by the data types themselves (each data type is associated with an ID). The APIs associated with this module are as follows

- initialize(registration_handler): initializations of the module's global variables, and the *registration_handler* callback are done here. *registration_handler* is a callback function called when a data registration packet is received. The configuration parameters are read in this function. This function maintains a key-value mapping of all registered sensors to their respective parsing format, indexed on a combination of *device_id*, *sensor_id* and *sensor_types*. The parsing format is obtained by parsing the sensor registration packet. A lookup in this dictionary will tell the Frame Parser how to parse each sensor data packet that it receives.
- parse_packet(device_id, data): Here, *device_id* refers to the ID of the sensor node, from which the packet is received. *data* is the sequence of bytes which make up the packet that is required to be parsed. The dictionary contains *sensor_id*, *sensor_type*, *rtc_timestamp* and *data* keys mapped to their respective values. The data field is again a dictionary, containing key-value pairs of *sen-*

sor_value_field—sensor readings. If the packet is a sensor registration packet, no value is returned.

The configuration parameters include *sensor_type*—value mappings and the data type associated with each data type ID. Here, data types refer to those defined by the Java standard. Defining unpacking type bindings for other types is also supported.

3.2 RF Driver

The main task of this module is to offload the data received from the sensor nodes, on to a receiver queue, which is then consumed by the Controller. Each sensor node is identified by a device ID, usually a unique ID given by the underlying protocol (IPv6 address in the case of 6LoWPAN, MAC Address/UUID for BLE). It is treated as a character sequence, for compatibility across protocols. The APIs associated with this module are as follows

- initialize_protocol(): The underlying protocol's initializations are carried out in this function. In the context of 6LoWPAN, initialization of the MAC and the PAN IDs, the creation of the 6LoWPAN network, etc., are performed here. The gateway acts as a 6LoWPAN border router [6] (LBR).
- send_payload(payload, device): *payload* should be sent to the sensor node identified by *device*. Payload is treated as a byte sequence.
- connect(receiver_queue, conf_path): the configuration parameters are read and the attributes are set accordingly. Initialization of the receiver queue to which the sensor data is dumped, is carried out here.
- list_connected_sensor_nodes(): returns the list of all currently active sensor node IDs.
- add_to_block_list(device): The sensor node identified by *device* is instructed to stop transmitting the sensor telemetry. If the protocol does not support such a feature, the values from that particular sensor node are ignored and not relayed back to the Controller.
- read_handler(): receives data from the communication port and maintains a buffer to accommodate partially received messages. Once a complete message is received, it is determined whether data message or a confirmation/error message has been received. Data messages are passed to the *get_payload*() function. Confirmation messages and error messages have protocol-specific handlers. The read_handler function is run on a separate thread.
- get_payload(): The data message received from read_handler() is parsed as per the protocol rules, and a dictionary containing *payload—device_id* mapping is created. This dictionary is then offloaded to the receiver queue.

The configuration parameters can include protocol-specific options.

3.3 Cloud Connectivity Client

Here, each sensor node is identified by a name instead of a device ID to make it more user-readable. The Controller is responsible for the conversion of device IDs to their respective device names. This is done so that the automation of device registration on the cloud under the same name and labeling the device in the cloud presentation layer is easier to implement. The user can provide the *device_id–device_name* mappings through the Dashboard. For multiple cloud platforms to be supported, each cloud platform must implement the following interfaces

- send_message(device_name, payload): push the payload to the cloud endpoint URL (defined as a configuration parameter) as JSON objects.
- register_event_callback(device_name, event_name, callback): this function allows an event handler to be assigned to an event, which is to be called when an incoming message arrives from the cloud, matching the given *event_description* (i.e., specific to a sensor node). Functions to forward the incoming message from the cloud to the respective sensor node can be added here (routing it through the Controller).
- register_device(device_name): register the new device as *device_name* on the cloud. This function is called by the Controller through a callback when a new sensor node joins the network.
- start_service(device_name): enable cloud service on a per-sensor node basis.
- initialize(default_message_callback): can be used to initialize the global variables, define various message callback handlers for each event and read the configuration files. A default message handler can also be specified here.

3.4 Controller

This module initializes all the other modules. It reads packets sent by the RF Driver and hands them over to the Frame Parser to be parsed. The parsed data is then sent to the Database, and to the Edge-Logic module which outputs the processed data. This data is forwarded to the cloud. The following are the APIs under this module.

- packet_registration_handler(device_id, parsed_registration_packet): this function is called by the Frame Parser through a callback when a sensor registration packet is received. The Database module is consulted to identify whether a new sensor node has joined, a new sensor type is being registered, or a new sensor of an existing type is being registered. The parsed information is sent to the database to create new tables, and to the Dashboard to display the newly added component. In the case of adding a new sensor node, the *register_device*() function of the Cloud Connectivity Client module is called. If an entry already exists for the sensor with the same name, type and ID, it is assumed that the sensor node has reconnected, after disconnecting.
- rq_reader(receiver_queue): Packets are read off the receiver queue one at a time and are passed on to the Frame Parser module for parsing. Data is sent to the

Database module to be added as a new record and to the Dashboard to be plotted as a new data point. It is sent to the Edge-Logic module to be processed, and the modified data is sent to the Cloud Connectivity Client's *send_message*() function.
- initialize(regs_ui_callback, data_ui_callback): the arguments of this function are callbacks to the Dashboard module's functions, that are to be called when a registration packet and a data packet are received. This function calls the *initialize*() functions of all other modules. It hands over the receiver queue handler to the RF Driver, and starts the thread to listen to packets from sensor nodes.

3.5 Dashboard

Various web application routes are defined in this module, both for serving the webpages as well as to poll sensor data and console messages from the client browser.

- get_msg(): This function servers the console messages to the web interface on the dashboard. This can be done by redirecting the stdout and stderr outputs using a custom tee() function.
- get_data(data_queues, device_name, sensor_type and sensor_id): The data_queues parameter is a dictionary of dictionary of dictionaries, following the device_name → sensor_type → sensor_id hierarchy. This variable maintains the latest value for each sensor, of each of the sensor types, belonging to each of the sensor nodes. Based on the given parameters, the latest value of that particular sensor is sent to the client.
- set_metadata(): this is the handler to the Controller callback that is called when a new sensor is registered with the Gateway. This updates the UI to include information about the newly connected sensor.
- update_options(): This function enables to configuration parameters of all the modules to be set from the Dashboard. The options are sent as JSON objects, which are in turn parsed and updated in each module.
- reinitialize_controller(): This function is registered as an event handler when the client requests to restart the Gateway. The Controller reinitializes all other modules.

3.6 Database

This module creates and maintains a database to store and query sensor data. In addition, it is used to store metadata about the sensors that are registered with the Gateway. We provide a template of the database table structures

- Metadata table
  - Id—a hash of the string "*device_name.sensor_type.sensor_id*" to uniquely identify a sensor.

  - device_name—name given to each sensor node by the user (in Controller configurations)
  - device_id—the id of the device (IPv6 for 6LoWPAN)
  - sensor_type
  - sensor_id—the id assigned for each sensor in the sensor node.
  - sensor_value_fields—the fields in the data measured by the sensor. For example, an accelerometer sensor measures x-axis, y-axis and z-axis sensor value fields.

- Sensor data table(s): One table for each sensor type is created, as the number of *sensor_value_fields* may vary for each sensor type. The *sensor_type* string is used as the table name. Data from all sensors (across sensor nodes) of the same *sensor_type* are added to the same table. The columns defined in this table are:

  - id
  - rtc_timestamp
  - sensor_value_fields[0]
  - sensor_value_fields [1]
  - sensor_value_fields [3]
  - …

3.7 Edge-Logic

This module can be customized based on the application the sensors are being used for. Since the sensor node metadata is forwarded along with the sensor data, the edge-logic operations can be sensor node-specific.

The custom code provided must implement the following interface. The code can also define custom functions to perform persistent background tasks (to mimic cloud worker roles) can be run on separate threads.

- initialize(db_handler): any initializations such as creating threads and global variable values, pertaining to the edge-logic app can be done here. The database handler is received in this function.
- stream_data(device_name, data, timestamp): Here, the edge-analytics is perform on streaming data packets sent by the sensor nodes. A dictionary containing processed data that needs to be sent to the cloud is returned by this function. It is invoked directly by the Controller.

4 Implementation Specifications

The project goals include code readability (and thereby more maintainability) and fast development lifecycles as it is required to be quickly and easily scalable in terms of feature support. Python3 was found to be an ideal choice for the programming

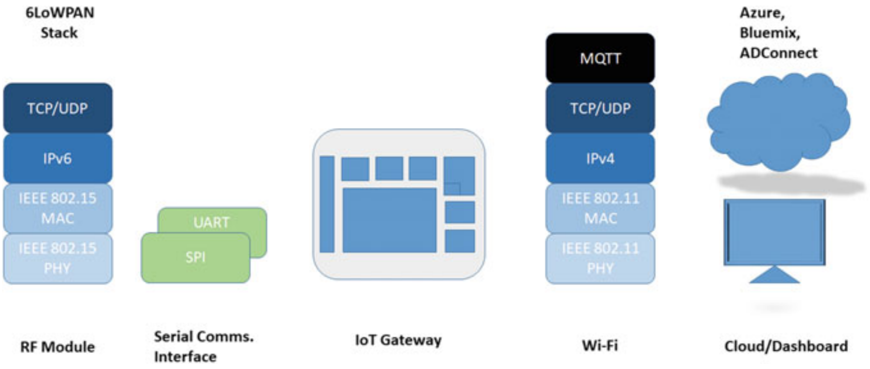


Fig. 2 Protocol stack of the IoT Gateway system—with the 6LoWPAN implementation

Table 1 Implementation specifications

Programming language	Python3.4
Database server	MySQL
Cloud communication protocol	MQTT
Web framework	Flask
Configuration file format	JSON

language, as ‘The Zen of Python’ [7] aligns with our goals. Most of the popular cloud platforms support their SDKs in Python.

The configuration parameters for each module was fed through JSON objects. A web interface was built on top of this using the Flask web framework which hosted the Dashboard. The web interface also includes the Chart.js graphing script which is used to plot the sensor values in real time. A MySQL database was used to store the sensor telemetry, and the MQTT cloud connectivity protocol was employed to communicate with the cloud as it brings high QoS and reliability to its applications [8] as compared to other protocols which also use the publish/subscribe pattern. Azure and ADConnect were our cloud platforms choice. Azure’s Python SDKs provided device management and security on Azure’s IoTHub. A sample cloud dashboard was developed using Azure’s Web App development service for the cloud presentation layer. The complete protocol stack for our PoC implementation is shown in Fig. 2. The implementation specifications are summarized in Table 1.

The entire package has been built to run on both Windows and Linux Systems. It was extensively tested on the Raspberry Pi 3, running Raspbian Jessie. In addition, its cross-platform characteristic allows other devices such as laptops to act as IoT Gateway systems while running our code. This is especially useful for testing and validating newly developed sensors. The entire software was hosted on a Git repository and an installation script was developed to facilitate easy installation and OTA update support. In addition, a sensor node simulator was developed as part of

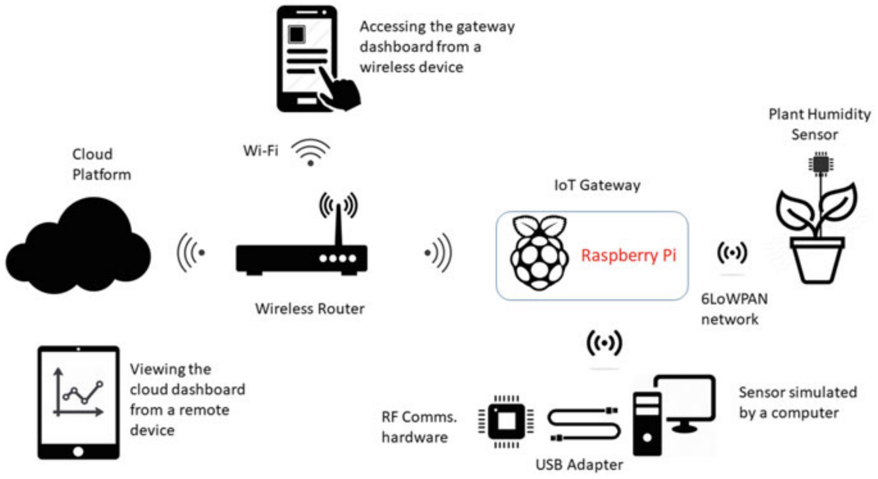

Accessing the gateway dashboard from a wireless device
Plant Humidity Sensor
Cloud Platform
Wi-Fi
IoT Gateway
Raspberry Pi
Wireless Router
6LoWPAN network
Viewing the cloud dashboard from a remote device
RF Comms. hardware
Sensor simulated by a computer
USB Adapter

Fig. 3 The IoT system layout—depicting the gateway, two sensor nodes, dashboard access, the cloud and a remote cloud dashboard viewer

this project, designed to periodically generate dummy 'sensor values' off a Windows PC/Linux box.

Figure 3 shows the layout of the IoT system. The IoT Gateway hosted on a Raspberry Pi 3 microcontroller receives sensor data through the 6LoWPAN network from a plant humidity sensor and a simulated light, accelerometer sensor node on a PC. The simulated sensor node communicates on the 6LoWPAN network using RF hardware connected to the PC using a USB adapter. On the other end of the spectrum, the Gateway is wirelessly connected to the Azure cloud through a router. The Gateway's Dashboard is viewed on a phone/tablet connected to the Gateway via the local network. The cloud dashboard is viewed from a remote device.

5 Results and Future Work

Our design of the IoT Gateway system was adopted for Structural Health Monitoring by the IoT Applications Team at Analog Devices, and was successfully presented at The IoT Solutions World Congress 2017, Barcelona. The team was able to extend the initial implementation with ease, to support the ADConnect cloud platform. The extension of this work to incorporate the SmartMesh technology is in progress.

The architecture and the design presented in this paper enables the development of cross-platform, multi-protocol smart IoT Gateway systems. Furthermore, it supports cloud platform compatibility and fast horizontal scalability across protocols. It allows the incorporation of custom application-specific edge-logic in the Gateway.

Acknowledgements We express our acknowledgements to Abhishek Kumar and Pranit Jadhav from Analog Devices India Pvt. Ltd., for their work in adopting the Gateway software to Structural Health Monitoring application.

References

1. Aazam, M., Khan, I., Alsaffar, A.A., Huh, E.N.: Cloud of things: integrating internet of things and cloud computing and the issues involved. In: 2014 11th International Bhurban Conference on Applied Sciences and Technology (IBCAST), pp. 414–419. IEEE, New York (2014)
2. Sun, W., Choi, M., Choi, S.: IEEE 802.11 ah: A long range 802.11 WLAN at sub 1 GHz. J. ICT Stand. **1**(1), 83–108
3. Jung, M., Weidinger, J., Reinisch, C., Kastner, W., Crettaz, C., Olivieri, A., Bocchi, Y.: A transparent IPv6 multi-protocol gateway to integrate building automation systems in the internet of things. In: 2012 IEEE International Conference Green Computing and Communications (GreenCom), pp. 225–233. IEEE, New York (2012)
4. Aazam, M., Huh, E.N.: Fog computing and smart gateway based communication for cloud of things. In: 2014 IEEE International Conference on Future Internet of Things and Cloud (FiCloud), pp. 464–470. IEEE, New York (2014)
5. Fog Computing and the Internet of Things: Extend the Cloud to Where the Things Are White Paper. https://www.cisco.com/c/dam/en_us/solutions/trends/iot/docs/computing-overview.pdf. Last accessed 9 Oct 2017
6. Olsson, J.: 6LoWPAN demystified. Texas Instrum. p. 13 (2014)
7. PEP 20—The Zen of Python. https://www.python.org/dev/peps/pep-0020/#id3. Last accessed 10 Oct 2017
8. Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., Ayyash, M.: Internet of things: a survey on enabling technologies, protocols, and applications. In: 2015 IEEE Communications Surveys & Tutorials, pp. 2347–2376. IEEE, New York (2015)

Privacy Preserving in Data Stream Mining Using Statistical Learning Methods for Building Ensemble Classifier

P. Chandrakanth and M. S. Anbarasi

Abstract As data evolves into streams, preserving privacy is a challenging task. Defining a classifier and understanding the categories of data on a dynamic data is an algorithmic task. Many algorithms have recommended using ensemble methods. Statistical ensemble methods are used in this paper to confine a considerable revision of the classifier and recommend the reliability of collecting data for preserving privacy. We conducted experiments on synthetic data and real-time data, and drawn algorithms to identify the drifts and recommendations to the classifier. The framework is experimented thoroughly and results are drawn.

1 Introduction

Advent of hardware technologies facilitates ability of using vast range electronic gadgets and data collection units, enable simple transactions on electronic terminals, Internet, and phone. Automation of data storage collects every data on piles and vaults. Data mining has a great collection of computational methods and algorithms [1, 2] to work on knowledge extraction. Researchers have attempted to work on the static data and dynamic data. Static data are the data sets and the structured databases; dynamic data includes change of the database structure and types used to definition. One of the dynamic models of data sets is data streams [3–5]. Working on data streams is nevertheless a challenging job for researchers having experience with large data sets. Such huge collected data invites computational challenges with some underlying facts: (a) it is not possible to fetch the accurate results in data processing, (b) in many cases; the daily data consists of temporal components, as data evolve over time, where adapting a one-pass mining algorithms will not give effective solutions, algorithms on stream mining are essential [6, 7]. The problems rise while change

P. Chandrakanth (✉) · M. S. Anbarasi
Pondicherry Engineering College, Puducherry, India
e-mail: chandrakanth@pec.edu

M. S. Anbarasi
e-mail: anbarasims@pec.edu

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_66

of data characteristics, drawing inferences properly from data streams. Change of characteristics symbolically or semantically leads to concept drift [7, 4].

1.1 Concept Drift in Data Streams

The concept drift means that the statistical properties of the target variable [8], which the model is trying to predict, change over time in unforeseen ways. Sudden change of concept in the data stream is not usual, but the change of values that has different statistical qualities from one time window to other. The Changing of concepts is concept drift, the classifier becomes a wrong model to apply on the data with drifts and thus classification becomes unpredictive. To overcome the unpredictive nature of classifiers, this opportunistically enhances classifiers in the classification process. The revision of classifiers [9] leads to derivation of new models. A concept drift stream [10, 7, 4] is observed by examining the changing statistical qualities in data [8] from each time window, or the nature of data and its characteristics with respect to the domain.

Ensemble classification is widely used using machine learning methods which was found to be more reliable. Learning methods infest much logic and are specific to each domain which will required domain knowledge and decision trees are built for each set of arising attributes. The sense of detecting the rising attributes are not clearly mentioned in these algorithms, except the domain expert pays for the application.

Ensemble classification is also applied using dimensionality reduction, overcoming the curse of dimensionality, reducing the attribute selection risks and identifying the potential attributes that has personally identifiable information with respect to the privacy preserving data mining [11, 12].

In our proposed work we apply the concepts of statistical learning and identify the arising attributes in the streams by calculating the probability density function on the univariate and multivariate data sets in a given time window.

The probability density function determines the level of inferences arrives at calculating the mean and variance of the data sets. The statistical properties remain same if there is no arrival of new attribute for a given time window. If these properties vary from each time window then there is a chance of arrival of a new attribute and then signals the generation of ensemble classifier. This is a most important task that would made comfortable in every stream processing application if the new attribute arrival and generation of ensemble classifier is at ease.

2 A Framework and Algorithm for Ensemble Classification

Let $\boldsymbol{w}_1$, $\boldsymbol{w}_2$ and $\boldsymbol{w}_3$ be the windows in the data stream allotted with different (though timestamp is not recorded in the stream) timestamps $\boldsymbol{t}_1$s and $\boldsymbol{t}_1$e, $\boldsymbol{t}_2$s and $\boldsymbol{t}_2$e, and $\boldsymbol{t}_3$s

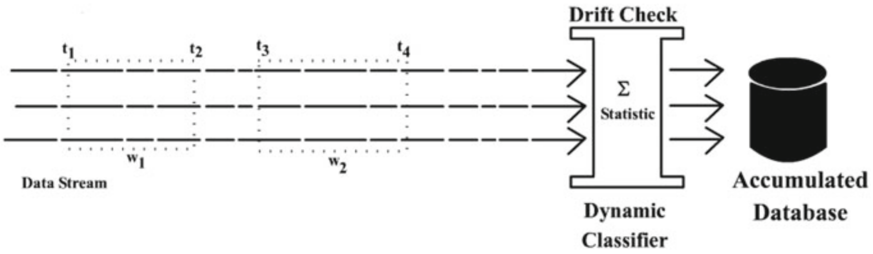


Fig. 1 Building dynamic classifier from data streams using drift checking

and t_3e are the start and end times of each window respectively. Let the microstate of each window is denoted by μ. μ_1 is the microstate of the first sample, μ_2 is the microstate of the second sample and μ_n is the microstate of the nth sample. $\mu_i w_j$ is the ith microstate calculated for jth window. The collection of microstates in each window is a continuous random variable and the microstate is the mean of percentages of instance names (category) in a two-input-table. As uncountable samples can be made on the window, these microstates on samples of window form a continuous random variable that takes on an infinite number of possible values. Assuming a support threshold S, the finite number of microstates is selected for calculating the probability density function, which the values of microstates from samples of window become finite (Fig. 1).

3 Datasets

Data sets for algorithmic experimentation are taken in two fold. First, we have tested the framework with a synthetic data stream simulator considering the hospital domain. Data generation for algorithmic potential test, propensity of algorithm is an important task for algorithm-based works. The data available as real time may not be directly suitable for processing with the data mining algorithms, converting the data into suitably formed, noise eliminated data sets is very essential. By collecting the characteristics of the real-time data related to medical and health care areas, data is synthetically generated for checking the potential and propensity of the algorithm. Synthetic data generation is the key process in the tool developed for the accomplishing the solution of the research. Second, to test the intense of algorithms we have selected real-time data from UCI Web Data Repository. The data set related Primate splice-junction gene sequences (DNA) with associated imperfect domain theory are selected from UCI Web Data Repository.

4 Experimental Results

The results shown in this section belong to the experimentation on the real-time data sets extracted from UCI Web Data Repository. A statistical method for learning and building the ensemble classifier is adopted in the framework. The data is supplied by a threaded stream generator and the time windows are adjusted during experimentation. The pause of stream is used to analyze the nature of data in the time window to calculated the variance and deduce the microstates. The probability density function is calculated on the microstates. For the experimentation a microstate is a mean of the variable in the stream pertaining to the time window. The microstates are further used to calculate the probability density function and subsequently checked for each time window to discover the drift. The changed microstates are supplied to the generation of classifier, the classifier contains all the attributes that are used with a threshold of microstates, the change of attributes in the collection of ensemble occurs when there is a change in the microstate and subsequently in the probability density function.

4.1 Methods

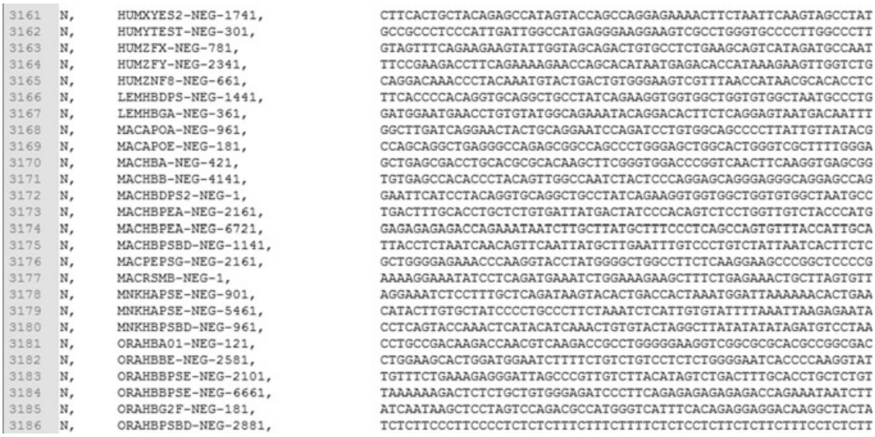
We apply in this work, statistical mechanics to derive microstates and probability density function based statistical learning mechanism. The pdf is calculated on the microstates and the input is given to the classifier recommender. The classifier recommender works with set relational minus operation and discovers and attributes new changes to the criterion in the classifier. Thus building a dynamic classifier or ensemble classifier.

Downright factors speak to sorts of information which might be partitioned into gatherings. Cases of absolute factors are showing class, occurrence names, and arrangements. While the last two factors may likewise be considered in a numerical way by utilizing precise esteems for example names and arrangements, it is frequently more useful to classify such factors into a moderately modest number of gatherings (Fig. 2; Table 1).

Investigation of downright information for the most part includes the utilization of information tables. A two-way table presents unmitigated information by checking the quantity of perceptions that fall into each gathering for two factors, one isolated into lines and the other separated into sections.

A two-way table representing EI (Entron, Intron) class with a prefix trimmed instance name and their counts calculated for microstate.

Since basic checks are frequently hard to dissect, two-way tables are regularly changed over into rates. In the above case, there are 6 tuples with an Instance Class HUM. Since there were a total of 24 observations, this means that 25% of the tuples belong to Instance Class HUM. If investigation with percentages is required, within a given category—of the 6 tuples of Instance Class HUM, 2 (33.3%) have the nucleotides AT and GC, 1 (16.55%) has nucleotides TG and CA. The microstate



```
N,      HUMXYES2-NEG-1741,              CTTCACTGCTACAGAGCCATAGTACCAGCCAGGAGAAAACTTCTAATTCAAGTAGCCTAT
N,      HUMYTEST-NEG-301,               GCCGCCCTCCCATTGATTGGCCATGAGGGAAGGAAGTCGCCTGGGTGCCCCTTGGCCCTT
N,      HUMZFX-NEG-781,                 GTAGTTTCAGAAGAAGTATTGGTAGCAGACTGTGCCTCTGAAGCAGTCATAGATGCCAAT
N,      HUMZFY-NEG-2341,                TTCCGAAGACCTTCAGAAAAGAACCAGCACATAATGAGACACCATAAAGAAGTTGGTCTG
N,      HUMZNF8-NEG-661,                CAGGACAAACCCTACAAATGTACTGACTGTGGGAAGTCGTTTAACCATAACGCACACCTC
N,      LEMHBDPS-NEG-1441,              TTCACCCCACAGGTGCAGGCTGCCTATCAGAAGGTGGTGGCTGGTGTGGCTAATGCCCTG
N,      LEMHBGA-NEG-361,                GATGGAATGAACCTGTGTATGGCAGAAATACAGGACACTTCTCAGGAGTAATGACAATTT
N,      MACAPOA-NEG-961,                GGCTTGATCAGGAACTACTGCAGGAATCCAGATCCTGTGGCAGCCCCTTATTGTTATACG
N,      MACAPOE-NEG-181,                CCAGCAGGCTGAGGGCCAGAGCGGCCAGCCCTGGGAGCTGGCACTGGGTCGCTTTTGGGA
N,      MACHBA-NEG-421,                 GCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCCGGTCAACTTCAAGGTGAGCGG
N,      MACHBB-NEG-4141,                TGTGAGCCACACCCTACAGTTGGCCAATCTACTCCCAGGAGCAGGGAGGGCAGGAGCCAG
N,      MACHBDPS2-NEG-1,                GAATTCATCCTACAGGTGCAGGCTGCCTATCAGAAGGTGGTGGCTGGTGTGGCTAATGCC
N,      MACHBPEA-NEG-2161,              TGACTTTGCACCTGCTCTGTGATTATGACTATCCCACAGTCTCCTGGTTGTCTACCCATG
N,      MACHBPEA-NEG-6721,              GAGAGAGAGACCAGAAATAATCTTGCTTATGCTTTCCCTCAGCCAGTGTTTACCATTGCA
N,      MACHBPSBD-NEG-1141,             TTACCTCTAATCAACAGTTCAATTATGCTTGAATTTGTCCCTGTCTATTAATCACTTCTC
N,      MACPEPSG-NEG-2161,              GCTGGGGAGAAACCCAAGGTACCTATGGGGCTGGCCTTCTCAAGGAAGCCCGGCTCCCCG
N,      MACRSMB-NEG-1,                  AAAAGGAAATATCCTCAGATGAAATCTGGAAAGAAGCTTTCTGAGAAACTGCTTAGTGTT
N,      MNKHAPSE-NEG-901,               AGGAAATCTCCTTTGCTCAGATAAGTACACTGACCACTAAATGGATTAAAAAACACTGAA
N,      MNKHAPSE-NEG-5461,              CATACTTGTGCTATCCCCTGCCCTTCTAAATCTCATTGTGTATTTTAAATTAAGAGAATA
N,      MNKHBPSBD-NEG-961,              CCTCAGTACCAAACTCATACATCAAACTGTGTACTAGGCTTATATATATAGATGTCCTAA
N,      ORAHBA01-NEG-121,               CCTGCCGACAAGACCAACGTCAAGACCGCCTGGGGGAAGGTCGGCGCGCACGCCGGCGAC
N,      ORAHBBE-NEG-2581,               CTGGAAGCACTGGATGGAATCTTTTCTGTCTGTCCTCTCTGGGGAATCACCCCAAGGTAT
N,      ORAHBBPSE-NEG-2101,             TGTTTCTGAAAGAGGGATTAGCCCGTTGTCTTACATAGTCTGACTTTGCACCTGCTCTGT
N,      ORAHBBPSE-NEG-6661,             TAAAAAAGACTCTCTGCTGTGGGAGATCCCTTCAGAGAGAGAGAGACCAGAAATAATCTT
N,      ORAHBG2F-NEG-181,               ATCAATAAGCTCCTAGTCCAGACGCCATGGGTCATTTCACAGAGGAGGACAAGGCTACTA
N,      ORAHBPSBD-NEG-2881,             TCTCTTCCCTTCCCCTCTCTCTTTCTTTCTTTTCTCTCCTCTTCTCTTCTTTCCTCTCTT
```

Fig. 2 Snapshot of dataset

Table 1 Categorical data analysis for given data sets (partially) for the instance class EI, with derived instance name categories and sequence fragments

For the instance class: EI					
Instance name (category)	Sequence (fragments)				Total
	AT	TG	GC	CA	
HUM	2	1	2	1	6
BAB	1	1	0	2	4
CHP	1	0	0	4	5
GIB	1	0	4	1	6
GCR	1	0	2	0	3
Total	6	2	8	8	24

can be calculated with mean value of the totals of each nucleotide. (6 + 2 + 8 + 8)/4 = 6. The mean of microstates vary when there is an introduction of new Instance Name (Category) of the same instance class. When the percentages of total nucleotides of each Instance Name (Category) are calculated, the difference between the mean of microstates of two subsequent windows marks the drift (Fig. 3).

The mean of $\boldsymbol{i}$ microstates, the standard deviation and the microstate at which the evaluation is required determine the normal probability density function of jth window. For all $\boldsymbol{j}$ windows the normal probability density function is calculated.

Mean of Samples $(\boldsymbol{w_1}\ (\boldsymbol{t_1}\mathrm{s}, \boldsymbol{t_1}\mathrm{e})) = \boldsymbol{\mu_1 w_1}$
Mean of Samples $(\boldsymbol{w_2}\ (\boldsymbol{t_2}\mathrm{s}, \boldsymbol{t_2}\mathrm{e})) = \boldsymbol{\mu_2 w_2}$
Mean of Samples $(\boldsymbol{w_3}\ (\boldsymbol{t_3}\mathrm{s}, \boldsymbol{t_3}\mathrm{e})) = \boldsymbol{\mu_3 w_3}$

```
declare
 stream structure;
 ts,te   integer;
 attrs   integer;
 attrval   array;
 samples  integer;
 pdf      float;
begin
 read stream;
 set time window using ts and te // ts -> start time; te -> end time
 derive samples in window
 for all samples in window
  set a sample
  calculate individual attribute microstsates
  compute categorical 2-attribute tables
  calculate microstate for sample
 next
 compute mean for all samples
 compute standard deviation for all samples
 from the continuous random variable
 set threshold
 filter the states above threshold
 compute the normal probability density function // pdf
 store the pdf  // for future learning
end
```

Fig. 3 Algorithm for drift check and building the statistical corpus of probability densities of time windows

$$\begin{aligned}\text{p.d.f.}\left(\boldsymbol{w}_j\right) &= \text{normal distribution of } \left(\boldsymbol{\mu}_i\boldsymbol{w}_j, \text{Mean}\left(\boldsymbol{\mu}_i\boldsymbol{w}_j\right), \boldsymbol{\sigma}\left(\boldsymbol{\mu}_i\boldsymbol{w}_j\right)\right) = f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \\ &= \frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)_2} \qquad (1)\end{aligned}$$

The probability density function value of time windows is compared with a threshold difference and if the difference of comparison is above threshold, there is an identification of drift in the stream. Hence, even a minor drift even requires the classifier revision.

4.2 Performance

Classifier execution depends extraordinarily on the attributes of the information to be arranged. There is no single classifier that works best on every given issue (a wonder that might be clarified by the without no lunch hypothesis). Different experimental tests have been performed to contrast classifier execution and with discover the

Table 2 A confusion matrix to assess the nature of the results in the experiment

	Classes predicted		
Current classes		True class	False class
	True class	True positive	False positive
	False class	True negative	False negative

Table 3 The results of applying classification algorithms on UCI and synthetic data sets

Algorithms used	Data sets	K-nearest neighbor	Naïve Bayes	Statistical learning
Performance indicators				
Accuracy	UCI	88.64	77.14	92.54
	Synthetic	91.25	81.52	96.29
Error	UCI	11.36	22.86	7.46
	Synthetic	8.75	18.48	3.71
F-measure	UCI	0.8864	0.7714	0.9254
	Synthetic	0.9125	0.8152	0.9629

attributes of information that decide classifier execution. Deciding an appropriate classifier for a given issue is however still more a workmanship than a science.

The measures accuracy and review are prominent measurements used to assess the nature of an order framework.

The consistency of the experiment is evaluated from the analysis on the results obtained by the experiment. A confusion matrix is shown below to assess the nature of the results in the experiment (Table 2).

A measure that combines precision and sensitivity is the harmonic mean of the two parameters, that is called F-measure.

$$\text{F-measure} = \frac{2 \times \text{tp} \times \text{precision}}{\text{tp} + \text{precision}} \tag{2}$$

Exactness in the consequences of examination is measured as the extent of aggregate number of right forecasts ascertained as the proportion between the quantity of cases accurately ordered and add up to number of cases (Table 3).

$$\begin{gathered}\text{Accuracy} = \frac{\text{tp} + \text{fn}}{N}\\ \text{Error} = 1 - \text{Accuracy}\\ \text{Sensitivity} = \frac{\text{tp}}{\text{tp} + \text{fn}}\end{gathered} \tag{3}$$

The aftereffects of applying characterization calculations show that measurable learning techniques gives the most elevated exactness and effectively arranges

92.54% of cases in the UCI information display 3190 occurrences. It is watched that order calculations perform better on huge informational collections with countless.

5 Conclusions

Identifying the concept drift is a challenging task in data streams, as data evolves dynamically. We have developed a framework to identify the concept drift and recommendation to the classifier and filtering data sets for the privacy preserving mechanisms. Statistical ensemble methods drawn from statistical of mechanics are used in this paper to confine a considerable revision of the classifier and recommend the reliability of collecting data for preserving privacy. The concept of microstates and pdf values plays a key role to the dynamic revision of the classifier. Experimentation is performed using synthetic data and real-time data, and the framework is drawn into algorithms.

References

1. Aggarwal, C.C.: Privacy-Preserving Data Mining. Data Mining, pp. 663–693. Springer International Publishing, Basel (2015)
2. Bairagi, N.: A survey on privacy preserving data mining. Int. J. Adv. Res. Comput. Sci. **8**(5), 896 (2017)
3. Bifet, A., Kirkby, R.: Data Stream Mining a Practical Approach (2009)
4. Gao, J., et al.: A general framework for mining concept-drifting data streams with skewed distributions. In: Proceedings of the 2007 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics (2007)
5. Hulten, G., Spencer, L., Domingos, P.: Mining time-changing data streams. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York (2001)
6. Bifet, A.: Adaptive stream mining: pattern learning and mining from evolving data streams. In: Proceedings of the 2010 Conference on Adaptive Stream Mining: Pattern Learning and Mining from Evolving Data Streams. Ios Press, Amsterdam (2010)
7. Gaber, M.M., Zaslavsky, A., Krishnaswamy, S.: Data Stream Mining. Data Mining and Knowledge Discovery Handbook, pp. 759–787. Springer US (2009)
8. Santra, S.B.: A Notes on Thermodynamics and statistical mechanics. Department of Physics, IIT Guwahati (2014)
9. Alvear-Sandoval, R.F., Figueiras-Vidal, A.R.: On building ensembles of stacked denoising auto-encoding classifiers and their further improvement. Inf. Fusion **39**, 41–52 (2017)
10. Bose, R.P.J.C., et al.: Dealing with concept drifts in process mining. IEEE Trans. Neural Networks Learn. Syst. **25**(1), 154–171 (2014)
11. Dietterich, T.G.: Ensemble methods in machine learning. Multiple Classifier Syst. **1857**, 1–15 (2000)
12. Haixiang, G., et al.: Learning from class-imbalanced data: review of methods and applications. Expert Syst. Appl. **73**, 220–239 (2017)
13. Gama, J., et al.: A survey on concept drift adaptation. ACM Comput. Surv. (CSUR) **46**(4), 44 (2014)

A Novel Public Key Crypto System Based on Bernstein Polynomial on Galois Fields 2^m to Secure Data on CFDP

Smitha Sasi and L. Swarna Jyothi

Abstract The Consultative Committee for Space Data Systems (CCSDS) is a international institution of space agencies devoted on together designing regulated information handling strategies to support space investigation in addition to space science and functions. To facilitate the mass memory with extensive records, interoperability, system constellations, extended onboard use of real-time operating system CCSDS administers a flexile and proficient data transmission protocol is CCSDS File Delivery Protocol. CFDP supports the convenience of application layer in internet protocol stack. Data transaction over the layers takes place in two different mode, either unreliable or reliable. However, the main demand in the data communication network is the confidentiality of data. With respect to confidentiality, cryptanalysis is utilized to encrypt data occupying memory on storage devices or traversing through communication links to make sure that any illicit access is unsuccessful. Here in the paper suggests secured conveyance of CFDP information swapping over IP by CCSDS space information transaction, like, Telemetry and Telecommand in unproven approach of activity. Bernstein open key cryptographic calculation is employed for providing security. Public key cryptosystem also known as the asymmetric crypto system has increased protection in contrast to secret key technique a combination of linked keys are utilized by the sending and receiving entity. The issue that arises in most of the existing techniques the plain text is treated as an integer value which gives rise to decreased security. In this paper we have suggested a competent polynomial established public key cryptography method over Galois field, which regards plaint text as (x, y) coordinate points.

S. Sasi (✉)
VTU RRC, Belgavi, India
e-mail: smitha.sasi24@gmail.com

L. Swarna Jyothi
Department of Electronics and Communication Engineering, Rajarajeswari College of Engineering, Bangalore, India
e-mail: swarnajyothi57@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_67

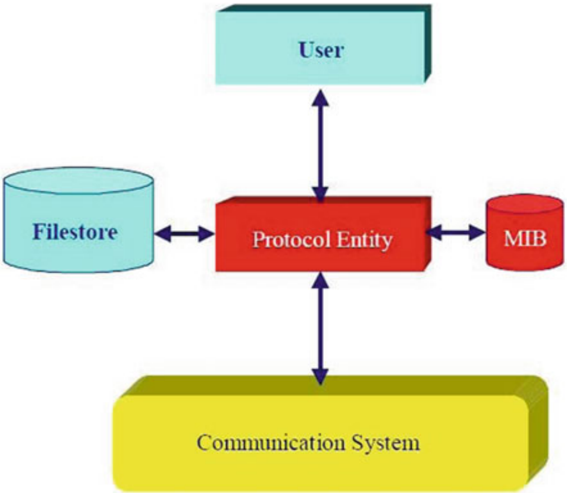
1 Introduction

The modern space data communication braces various space missions with extensive file transaction on space data channels. For supporting the effective data communication over the links, CCSDS advocates a novel protocol named as CCSDS File delivery protocol (CFDP) [1]. CFDP adopts the characteristics of the application layer and accomplishes both unreliable and reliable approaches. Though, the considerable demand in data communication is to cater the confidentiality over the information [2]. Cryptography is an compelling method to safeguard delicate data when it is stored in memory of any storage device or transmitted through network communication channel [3]. Cryptographic algorithms may be actualized either on symmetric key or asymmetric key frameworks. The research supports security in the data transmission over space data link data utilizing asymmetric key systems. The Bernstein public key algorithm over Galois field in which the plain text is denoted as coordinate elements as basis on polynomial is used in this research. The mathematical calculation has less complexity in Bernstein polynomial [4]. This research proposes confidential communication of CFDP information transactions over IP over CCSDS space information connections, like, Telemetry and Telecommand in debatable approach of operation [5, 6].

2 CCSDS File Delivery Protocol

CFDP caters the facility to transmits “documents” between a space craft and mass memory. Files are sent back and forth automatically or by manual means. It is possible to exchange files in all the transmission modes like simplex, half-duplex and full-duplex. Record exchange may be activated consequently or manually [7]. The protocol is competent enough to work over an expanded assortment of mission outlines, from comparatively uncomplicated low earth orbit spacecraft to complicated set of orbiters and landers through the help of various ground facilities and transmission channels [8]. The protocol is autonomous of the applied science utilized to mechanize stockpiling the documents and makes use of just a small number of basic file storage capabilities to be able to operate. CFDP transmits and receives files to a local storage device, which CFDP calls the “file store”. Files to be sent are reclaimed from the file store and acquired files are recorded to the file store. Figure 1 explains the architecture regarding CFDP protocol. Here the end user communicates with protocol entity via various service primitives. The user is a software task or a human being. Protocol entity is the protocol implementation part. MIB is the Management Information Base which contains the timer and address mapping fields [9]. The CFDP architecture is as shown in Fig. 1.

Fig. 1 CFDP architecture



3 Bernstein Polynomial

In consideration with the developmental field of mathematical examination, a Bernstein polynomial, which is an amalgamation of Bernstein polynomials. A mathematically reliable technique to work with the survey polynomials in Bernstein structure is de Casteljau's count [10].

The $(n+1)$ Bernstein degree polynomials having n as their degree are symbolized as

$$n(f, t) = nc_i t^i (1-t)^{n-i}$$

nc_i is a binomial coefficient.

The Bernstein premise polynomials bearing n as their degree, have a structural premise for the vector space πn of polynomials normally having an nth degree [7].

The Bernstein basis polynomials with n as their degree create the foundation for the vector space Πn of polynomials of degree not beyond n.

The coefficient nc_i attained from pascal's triangle. The exponent of the $(1-t)$th term reduces by one as i is incremented.

- The Bernstein polynomials having a degree of 1 are

$$B_{0,1}(t) = 1 - t$$
$$B_{1,1}(t) = t$$

Graph for linear Bernstein where $0<t<1$ as depicted in Fig. 2.

- Bernstein polynomials having degree as 2 are

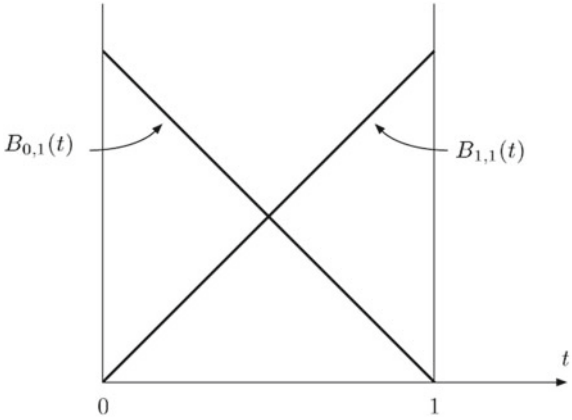


Fig. 2 Linear Bernstein polynomial

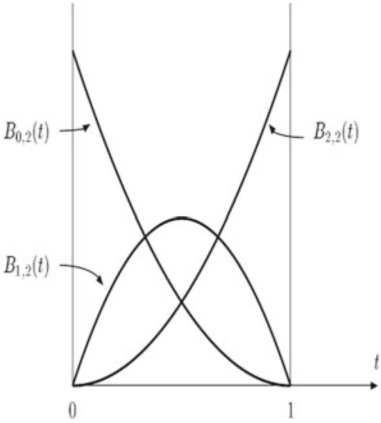


Fig. 3 Quadratic Bernstein polynomial

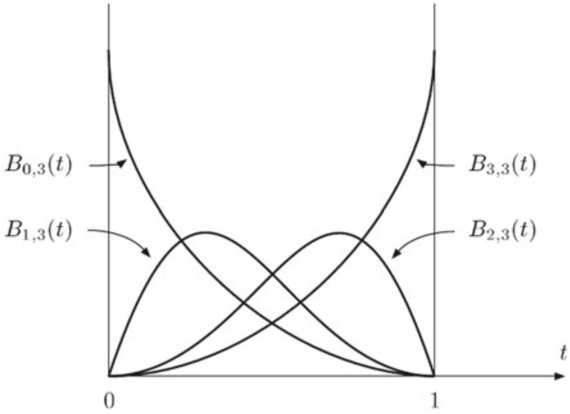
$$B_{0,2}(t) = (1-t)^2$$
$$B_{1,2}(t) = 2t(1-t)$$
$$B_{2,2}(t) = t^2$$

could be sketched for $0 < t < 1$ as in Fig. 3.

- Bernstein polynomials having degree as 3 are

$$B_{0,3}(t) = (1-t)^3$$
$$B_{1,3}(t) = 3t(1-t)^2$$
$$B_{2,3}(t) = 3t^2(1-t)$$

Fig. 4 Ternary Bernstein polynomial



$$B_{3,3}(t) = t^3$$

could be sketched for $0<t<1$ as displayed in Fig. 4.

4 Galois Field

There exists a finite field or Galois field (titled after Évariste Galois) in complex mathematics which comprises of a finite or a limited number of elements. The major applications for finite fields are in Galois theory, cryptanalysis, error control coding. A prime or power of a prime can be considered as order of finite field. Cryptanalysis targets the finite fields. The prime value should be larger than or equal to 1, there exists a exclusive field containing *pn* elements, depicted as GF(p^m). When n is equal to 1, the field is imod p, where $i = 1, 2, \ldots n$. In cryptography, when $p = 2$, then the field is 2, and it is termed as binary extension also denoted as GF(2^m) [11].

4.1 Arithmetic Operation Over Galois Field GF(2^{m})

Over GF(2^m), addition and subtraction modulo 2 processes are identical and this mathematical calculation could be carried out in hardware with the common XOR logic gate. The addition table for GF(2^3) is given in Table 1 and Multiplication table is given in Table 2 [12].

Table 1 Multiplication table over GF(2^3)

X	1	2	3	4	5	6	7
1	1	2	3	4	5	6	7
2	2	4	6	3	1	7	5
3	3	6	5	7	4	1	2
4	4	3	7	6	2	5	1
5	5	1	4	2	7	3	6
6	6	7	1	5	3	2	4
7	7	5	2	1	6	4	3

Table 2 Addition table over GF(2^3)

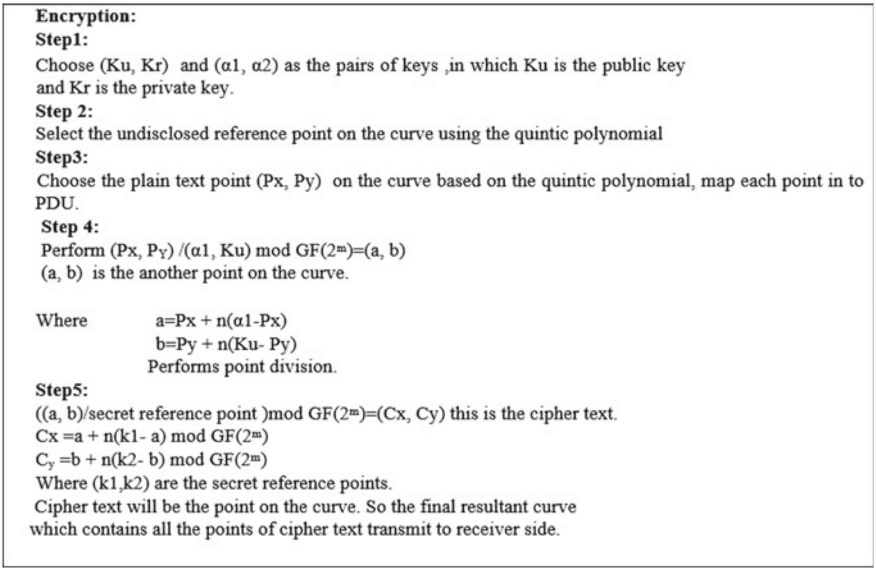
+	0	1	2	3	4	5	6	7
0	0	1	2	3	4	5	6	7
1	1	0	3	2	5	4	7	6
2	2	3	0	1	6	7	4	5
3	3	2	1	0	7	6	5	4
4	4	5	6	7	0	1	2	3
5	5	4	7	6	1	0	3	2
6	6	7	4	5	2	3	0	1
7	7	6	5	4	3	2	1	0

5 Proposed Algorithm

A familiar representation of the Bernstein polynomial is given as $n(f, t) = \sum_{r=0}^{n} f\left(\frac{i}{n}\right) nc_i t^i (1-t)^{n-i}$. When $n = 5$; the polynomial is a quintic polynomial.

Generate (x, y) points on the curve generated by Bernstein Polynomial. Map these points over the PDU [13].

The proposed algorithm implemented in socket programming in C [14].



Encryption:

Step1:

Choose (Ku, Kr) and ($\alpha1$, $\alpha2$) as the pairs of keys ,in which Ku is the public key and Kr is the private key.

Step 2:

Select the undisclosed reference point on the curve using the quintic polynomial

Step3:

Choose the plain text point (Px, Py) on the curve based on the quintic polynomial, map each point in to PDU.

Step 4:

Perform (Px, P_Y) /($\alpha1$, Ku) mod $GF(2^m)$=(a, b)

(a, b) is the another point on the curve.

Where $a=Px + n(\alpha1-Px)$

$b=Py + n(Ku- Py)$

Performs point division.

Step5:

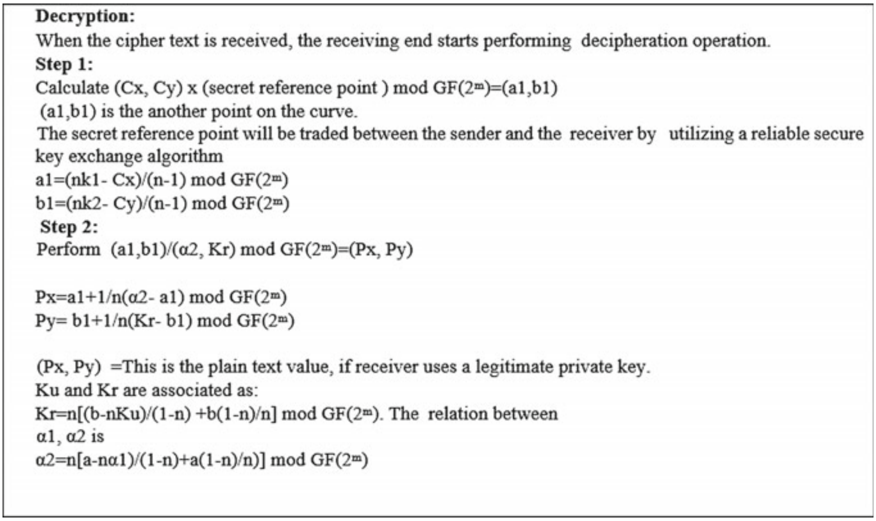
((a, b)/secret reference point)mod $GF(2^m)$=(Cx, Cy) this is the cipher text.

Cx =a + n(k1- a) mod $GF(2^m)$

C_y =b + n(k2- b) mod $GF(2^m)$

Where (k1,k2) are the secret reference points.

Cipher text will be the point on the curve. So the final resultant curve which contains all the points of cipher text transmit to receiver side.



Decryption:

When the cipher text is received, the receiving end starts performing decipheration operation.

Step 1:

Calculate (Cx, Cy) x (secret reference point) mod $GF(2^m)$=(a1,b1)

(a1,b1) is the another point on the curve.

The secret reference point will be traded between the sender and the receiver by utilizing a reliable secure key exchange algorithm

a1=(nk1- Cx)/(n-1) mod $GF(2^m)$

b1=(nk2- Cy)/(n-1) mod $GF(2^m)$

Step 2:

Perform (a1,b1)/($\alpha2$, Kr) mod $GF(2^m)$=(Px, Py)

Px=a1+1/n($\alpha2$- a1) mod $GF(2^m)$

Py= b1+1/n(Kr- b1) mod $GF(2^m)$

(Px, Py) =This is the plain text value, if receiver uses a legitimate private key.

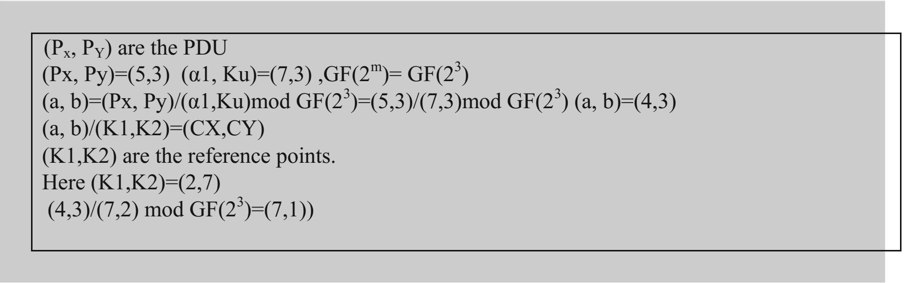
Ku and Kr are associated as:

Kr=n[(b-nKu)/(1-n) +b(1-n)/n] mod $GF(2^m)$. The relation between $\alpha1$, $\alpha2$ is

$\alpha2$=n[a-n$\alpha1$)/(1-n)+a(1-n)/n)] mod $GF(2^m)$

6 Result

Encryption:



(P_x, P_Y) are the PDU

(Px, Py)=(5,3) (α1, Ku)=(7,3) ,GF(2^m)= GF(2^3)

(a, b)=(Px, Py)/(α1,Ku)mod GF(2^3)=(5,3)/(7,3)mod GF(2^3) (a, b)=(4,3)

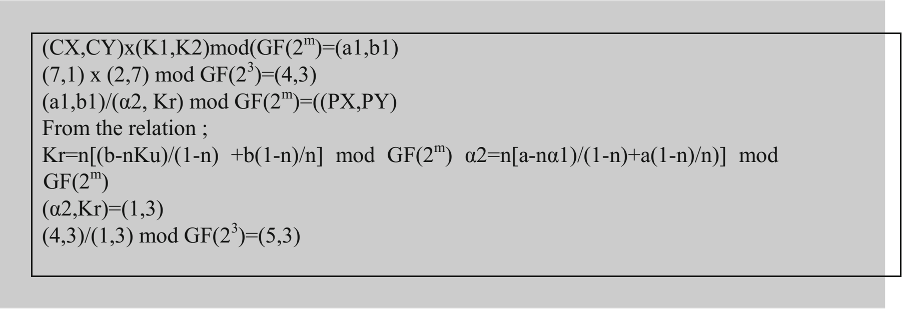
(a, b)/(K1,K2)=(CX,CY)

(K1,K2) are the reference points.

Here (K1,K2)=(2,7)

(4,3)/(7,2) mod GF(2^3)=(7,1))

Decryption:



(CX,CY)x(K1,K2)mod(GF(2^m)=(a1,b1)

(7,1) x (2,7) mod GF(2^3)=(4,3)

(a1,b1)/(α2, Kr) mod GF(2^m)=((PX,PY)

From the relation ;

Kr=n[(b-nKu)/(1-n) +b(1-n)/n] mod GF(2^m) α2=n[a-nα1)/(1-n)+a(1-n)/n)] mod GF(2^m)

(α2,Kr)=(1,3)

(4,3)/(1,3) mod GF(2^3)=(5,3)

Execution time for each public key crypto system analyzed shown in Table 3.

Table 3 Execution time analysis

m	RSA (s)	ECC	Bernstein (s)
6	3	2.039 s	1.03
4	1.03	1.06	0.06

7 Conclusion

Bernstein public key cryptosystem over Galois field is used for administering confidentiality on CFDP data communication. Here the keys are mathematically related public and private key pairs. This is the reason why only the individual having the legitimate decryption key can decipher an encrypted message. In comparison to other public key crypto methods this method provides added security in space data activity.

References

1. CCSDS File Delivery Protocol (CFDP). Recommendation for Space Data System Standards, CCSDS 727.0-B-3. Blue Book. Issue 3. Washington, D.C. (June 2005)
2. Space Data Link Security Protocol. Issue 1. Recommendation for Space Data System Standards (Blue Book), CCSDS 355.0-B-1. Washington, D.C. (September 2015)
3. Stallings, W.: Cryptography and Network Security, 3rd edn., pp. 42–62,121–144, 253–297. Pearson Education, ISBN 81-7758-011-6
4. Caglar, H., Akansu, A.N.: A generalized parametric PR-QMF design technique based on Bernstein polynomial approximation. IEEE Trans. Signal Process. **41**(7), 2314–2321 (1993)
5. TM Space Data Link Protocol. Recommendation for Space Data System Standards, CCSDS 132.0-B-1. Blue Book. Issue 1. Washington, D.C. (September 2003)
6. TC Space Data Link Protocol. Recommendation for Space Data System Standards, CCSDS 232.0-B-1. Blue Book. Issue 1. Washington, D.C. (September 2003)
7. Sanjay, P., Unnikrishnan, E., Lakshminarasimhan, P.: Design, implementation and performance evaluation of CCSDS CFDP Protocol. In: IEEE International Conference
8. IP over CCDS Recommendation for Space Data System Standards, CCSDS 702/1–R3. Red Book. Issue 1. Washington, D.C. (September 2003)
9. CCSDS File Delivery Protocol (CFDP). Introduction and Overview, CCSDS 720.1-G-3 Green Book. Issue 1 (April 2007)
10. Online Geometric Modelling Notes, Visualization and graphics research group. Department of Computer Science, University of California, Bernstein
11. Lin, S.: Error Control Coding, 2nd edn. Pearson Education. ISBN 978-0130426727
12. Sasi, S., Swarna Jyothi, L.: A heuristic cryptosystem based on bernstein polynomial on galois fields GF(P) and GF(2^m). IJLTEMAS **IV**(II) (February 2015)
13. Sasi, S., Swarna Jyothi, L.: A heuristic approach for secured transfer of image based on Bernstein Polynomial. In: IEEE International Conference on Circuits, Controls, Communication and Computing, Bangalore, pp. 312–315, ISBN: 978-1-4799-6546-5 (Nov 2014)
14. Stevens, R.: Unix Network Programming, 3rd edn. PHI, ISBN-13:978-0139498763

Compression of Medical Images Based on Devils Curve Coordinate System

B. Santhosh and Viswanath Kapinaiah

Abstract Medical Image has a great impact in health Industry, The storage and transmission are important challenges due to mass size of medical image information. There exists a need for compression of images for storage and transmission. To achieve high compression rate and less bandwidth for transmission, reliable and flexible compression technique required with loss of quality. This paper discusses a novel technique of compression for medical images based on polynomial based curve system. The properties of quartic curve supports digital image processing, One of the efficient quartic curve coordinates system is Devils Curve coordinate with the polynomial of degree 4. Proposed technique provides less computational complexity and higher compression ratio compared to conventional compression techniques.

1 Introduction

Medicinal imaging greatly affects determination of health issues and planning to surgery. The capacity and transmission is an essential difficulty because of colossal size of medicinal picture information. Current compression procedures give high compression rate however with loss of quality. Image compression is the technique reduces the size of the data required to display image digitally. Large amount of data is required to represent uncompressed image.

Image compression helps to reduce redundancy, increase transmission bandwidth, and reduce storage capacity. Compression can be classified as two techniques. Lossless compression—no information get lose during transmission, Lossy compression—this allows less than perfect reconstruction of the original image. The lossy is

B. Santhosh (✉) · V. Kapinaiah
Department of Telecommunication Engineering, Siddaganga Institute of Technology, SIT, Tumkur, India
e-mail: santhoshmehtre@gmail.com

V. Kapinaiah
e-mail: kviitkgp@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_68

more advantageous than lossless since less information is required. There are three main sources of redundant information [1] they are as follows:

Coding Redundancy–binary code used to represent grey values,

Inter pixel Redundancy–correlation between adjacent pixels in an image Psycho-**visual Redundancy**–unequal sensitivity of the human eye to different visual information.

In this paper, we discussed about the lossless compression technique by using (x, y) coordinate system. Here devils curve coordinate system used for compressing the image. This method is different from the conventional transform domain technique.

2 Methods for Compression

Image compression techniques can be categorizes as

(i) Lossless compression (reversible)
(ii) Lossy compression (irreversible).

Run-length encoded (RLE) and the JPEG techniques are lossless compression algorithms. A few data get discarded in lossy compression techniques. Lossy compression technique has high compression rate compared to lossless.

2.1 Lossless

In the technique of Lossless compression with the compressing of data that is when get decompressed, will be the same replica of actual data. In this case, when the binary data like the documents, executable, etc., are get compressed. This required to be reproduced exactly when get decompressed again. On the contrary, the images and the music also required not to be generated “exactly”. A resemblance of the actual image is sufficient for the most objective, as far as the error or problems between the actual and compressed image is avoidable or tolerable. These types of compression are also known as noiseless as they never add noise to signal or image. It is also termed as the entropy coding as it uses the techniques of decomposition/statistics to remove/reduce the redundancy. It is also used only for the some specific applications along with the rigid needs like a medical-imaging. The techniques for lossless compression are as follows:

- Huffman encoding
- Run-length encoding
- Arithmetic coding
- Dictionary Techniques.

2.2 Lossy

In the technique of lossy compression, it decreases the bits by recognizing the information which is not required by discarding it. Compression refers to minimizing the size of the data before storing or transmitting the data [2]. Dropping non-essential information from the source of data can save the storage area. As there are many variations in color in environment human eye will not differentiate changes in these variations. Lossy compression are widely used in modern-day cameras, mobiles, video codec, etc. In audio compression the psycho acoustics have been used to eliminate the inaudible signals [3].

3 Measures of Image Quality

Image quality is an important characteristic image processing to determine the image degradation. Mean-squared Error and Peak Signal-to-Noise Ratio values will decide the quality of image as shown in Eq. (2). The Mean-Squared Error (MSE) between two images are explained in Eq. (1):

$$\frac{1}{m \times n} \sum_{0}^{m} \sum_{0}^{n} (f - g)^2, \tag{1}$$

where

f matric of data image.
g matrix data of degraded image.
m number of rows of pixels of the images.
i index of that row.
n number of columns of pixels of the image.
j index of that column.

$$\mathrm{PSNR} = 20 \log_{10} \; \mathrm{Max}_f / \mathrm{sqrt(MSE)} \tag{2}$$

Max_f is the maximum signal value.

4 Existing Methods for Compression Techniques

4.1 DCT

The DCT transforms is a technique which transforms image from spatial domain to frequency domain [4, 5]. DCT helps to divide the image into parts based on image visual quality. The encoding equation for 1D DCT is given in (3).

DCT Encoding 1D DCT:

$$\frac{2}{N}^{1/2}\sum_{0}^{N}\wedge(i)\cos\left[\frac{\pi\times x}{2\times N}(2-i)\right]\times f(i) \tag{3}$$

Equation (4) is Inverse 1D DCT given as below:

$$\wedge(i)=\left\{\frac{1}{\sqrt{2}}\xi=0,1\text{ otherwise}\right. \tag{4}$$

The DCT encoding techniques is as shown in Fig. 1.

The encoding equation for 2D DCT is given in (5):

$$F(x,y)=\frac{2}{N}^{1/2}\frac{2}{M}^{1/2}\sum_{0}^{N}\sum_{0}^{M}\wedge(i)\cos\left[\frac{\pi\times x}{2\times N}(2-i)\right]$$
$$\times\cos\left[\frac{\pi\times y}{2\times N}(2-i)\right]\times f(i) \tag{5}$$

Inverse 2D DCT transform is shown in Eq. (6) and (7). Where

$$\wedge(i)=\left\{\frac{1}{\sqrt{2}}\xi=0,1\quad\text{otherwise}\right. \tag{6}$$

The algorithm to obtain DCT is as follows:

Input image is $N\times M$; $f(i,j)$ intensity of the pixel. $F(x,y)$ is the DCT coefficient of the DCT matrix. Upper left corner of the DCT specifies signal energy coefficients, signal energy lies at low frequencies. Higher frequencies in the lower right values

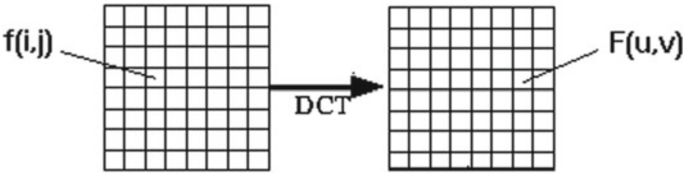


Fig. 1 DCT encoding technique

are small to neglect and is used for compression. DCT input is an 8 × 8 array, these pixel values vary from 0 to 255. Therefore an 8 point DCT would be: where

$$\wedge(i) = \begin{cases} \frac{1}{\sqrt{2}}\xi = 0{,}1 \text{ otherwise} \end{cases} \tag{7}$$

The DCT transform technique one of the efficient method in image compression. Transformation is orthogonal and fast algorithms can be used to compute and output consists of large number of zero values. Major disadvantage of the DCT if the input preprocessed 8 × 8 blocks are integer valued and the output values are real. So in quantization step values in each DCT block and output are integer valued.

4.2 DWT

The discrete wavelet transform (DWT) are small waves situated in various times. The waves are obtained by scaling and translation of a scaling function and wavelet function. This technique is limited in both time and frequency. Wavelet transform provides a multiresolution framework utilized in many applications. Wavelet can be built from a scaling function that must be orthogonal to its discrete translations [6, 7]. DWT decomposes the signal into mutually orthogonal set of wavelets, which is main difference from Continuous wavelet Transform (CWT) or its implementation for discrete time continuous wavelet transform (DT-CWT). The dilation equation is given in (8).

$$\phi(x) = \sum_{-\infty}^{\infty} (a_k \phi(S_x - k)) \tag{8}$$

DWT is one of the effective method for compression. But the major disadvantage of this method is greater complexity. The greater complexity translates into more resources requires more computation time as well as storage. The interpretation of the experimental result is complex and difficult.

5 Devils Curve Coordinate System

General Equation for Devils Curve is as shown in Eqs. (9) and (10):

$$y^2\left(y^2 - a^2\right) = x^2\left(x^2 - b^2\right) \tag{9}$$

Devil’s curve is named after the juggling game diabolo, which consists of two sticks, a string and a spinning prop in the likeness of the lemniscate [8].

Devil’s curve for $a = 0.8$ and $b = 1$. Equivalent is as shown in Fig. 2.

Fig. 2 Devil's curve

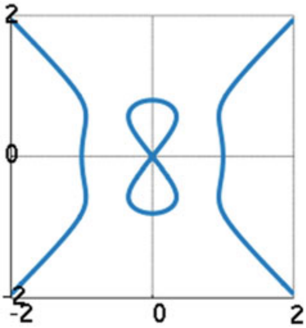
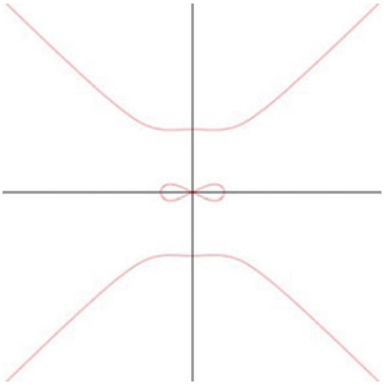


Fig. 3 Crunode at the origin



$$y^2\left(y^2 - a^2\right) = x^2\left(x^2 - b^2\right) \tag{10}$$

The polar equation is given in (11)

$$r^2\left(\sin^2\theta - \cos^2\theta\right) = a^2\sin^2\theta - b^2\cos^2\theta \tag{11}$$

And the parametric equations for x and y are given in (12) and (13)

$$x = \cos\, t\sqrt{\frac{a^2\sin^2 t - b^2\cos^2 t}{\sin^2 t - \cos^2 t}} \tag{12}$$

$$y = \sin\, t\sqrt{\frac{a^2\sin^2 t - b^2\cos^2 t}{\sin^2 t - \cos^2 t}} \tag{13}$$

The curve illustrated above corresponds to parameters $a^2 = 1$ and $b^2 = 1$.

It has a crunode at the origin for $a/b < 1$ as shown in Fig. 3, the central hourglass is horizontal, for $a/b > 1$, it is vertical, and as it passes through, the curve change to a circle a shown in Fig. 4 [9].

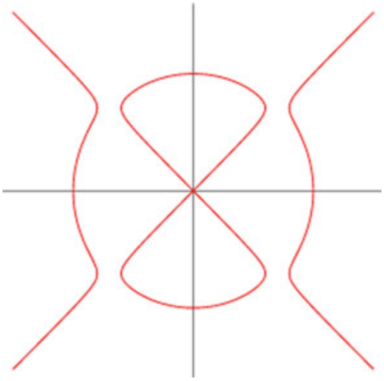


Fig. 4 Central hourglass horizontal

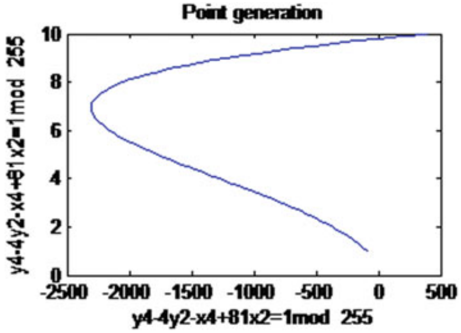


Fig. 5 Point generation

Image can be represented with more number of points, these points are interpolated to form different curves. Devil's curve and its applications in image processing have wide scope in research like pattern recognition, computer vision and computer aided geometric design. A novel cryptographic technique is used to provide confidentiality and authentication on medical images which is devil curve public key cryptosystem over prime field [10, 11].

Consider the devils curve polynomial, $y^4 - 4y^2 - X^4 + 81X^2 = 1 \mod 255$. Where the (x, y) points generated by the curve as shown in Fig. 5. The points on the curves are (13, 4), (40, 4), (5, 7), (48, 7), (6, 8), (47, 8), (25, 11), (28, 11), (2, 15), (17, 15) etc.

6 Proposed Algorithm for Compression

Procedure for compression starts by generating (x, y) points depends on the polynomial degree and obtain the image for the devils curve [12, 13].

Step 1: Consider DataImage is the input medical Image and DevilImage is the Image for devils curve based polynomials.

Table 1 Analysis of medical images with proposed algorithm

Proposed algorithm		
Image	PSNR (dB)	MSE
Ultrasound	44.30	2.24
X-Ray	45.38	3.78
CT scan	54.34	0.24

Step 2: The DataImage is divided by the image DevilImage is as shown below.

$$\frac{\text{DataImage}[x, y]}{\text{DevilImage}[x, y]} \bmod 255$$

CompressImage[x, y], obtained by point division formula [7, 14, 15].

$$\text{CompressImage} = \text{DataImage} + (\text{DevilImage} - \text{DataImage}) \times \bmod 255$$

Step 3: Transmit the compressed Images over Secure channel.
Step 4: Decompress the image recover the parent image.

$$\text{DecompressImage}[x, y] = \text{CompressImage}[x, y] \times \text{DevilImage}[x, y] \times \bmod 255$$
$$\text{DecompressImage}(x) = (\text{CompressImage}(x) - n \times \text{DevilImage}(x)/(1 - n)) \bmod 255$$
$$\text{DecompressImage}(y) = (\text{CompressImage}(y) - n \times \text{DevilImage}(y)/(1 - n)) \bmod 255$$

The range of n is $0<255$ (The prime numbers between this range).

7 Results and Discussion

In this approach, Compression technique is proposed and developed based on Devils curve coordinate system. The proposed idea is implemented using MATLAB. A set of test images are used to validate the efficient algorithm. The quality of medical image parameters like PSNR and MSE have been calculated and analyzed with existing techniques. Table 1 is an experimental analysis of proposed algorithm.

8 Conclusion

Medical image compression based on curve coordinate system is a novel approach in the medical image processing. In this approach, Devils curve coordinate compression technique used and experimentally tested with MATLAB tool. Devils curve over prime field approach is a lossless approach with PSNR ratio ranges from 24 to 32. This technique tested for various models with different size of data. The procedure followed in this were accurate and successful, there are always divergence with the experimental results. The reason for this discrepancies is due to unnoticeable errors will be accumulated from each stage.

References

1. Santhosh, B., Viswanath, K.: Review on medical image processing. In: Springer Conference in Advances in Intelligent Systems and Computing, vol. 435, Issue 1, pp. 531–537 (Jan 2016)
2. Cundy, H., Rollett, A.: Mathematical Models, 3rd edn, p. 71. Tarquin Publication, Stradbroke (1989)
3. Viswanath, K., Mukherjee, J., Mukhopadhyay, S., Pal, R.N.: Transcoding: JPEG2000 to JPEG. In: International Conference on Advanced Computing and Communication (ICACC), pp. 355–358 (Feb 2007)
4. Viswanath, K., Mukhopadhyay, J., Biswas, K.: Transcoding in the block DCT space. In: IEEE International Conference on Image Processing (ICIP) (2009)
5. Xiong, Z., Ramachandran, K., Orchard, M.T., Zhang, Y.-Q.: A comparative study of DCT and wavelet based image coding. IEEE Trans. Circuits Syst. Video Tech. **9**, 692–695 (1999)
6. Mallat, S.: A Wavelet Tour of Signal Processing. Academic Press, Cambridge, MA (1998)
7. Strang, G., Nguyen, T.: Wavelets and Filter Banks. Wellesley-Cambridge Press, Wellesley (1996)
8. Gray, A.: Modern Differential Geometry of Curves and Surfaces with Mathematics, 2nd edn, pp. 92–93. CRC Press, Boca Raton, FL (1997)
9. Curve co-ordinate system. http://www.2dcurves.com/quartic/quarticd.html
10. Philips: Interpolation and approximation of polynomials. G.M. ISBN: 978-0-38700215-6. http://www.Springer.Com/978-0-387-00215-6
11. Online Geometric Modelling Notes. Bernstein; Visualization and graphics research group; department of computer science. University of California. 15. Mathematical World. http://mathworld.wolfram.com/input.html
12. Santhosh, B., Viswanath, K.: The novel public key cryptosystem for medical image processing. In: IEEE Conference on Inventive Systems and Control ICISC 2017, pp. 49–50. ISBN 978-1-5090-4714-7 (Jan 2017)
13. Santhosh, B., Sasi, S.: Robustic technique to encrypt medical images using Bernstein polynomial over prime field. Int. J. Electr. Sci. Eng. **1**(1), 19–21 (2016). (ISSN: 2455-6068)
14. Sasi, S., Swarna Jyothi, L.: A heuristic approach for secured transfer of image based on Bernstein polynomial. In: IEEE International Conference on Circuits, Controls, Communication and Computing, pp. 312–315. ISBN: 978-1-4799-6546-5 (Nov 2014)
15. Cramer, G: Introduction a l'analyse des lignes courbes algbriques, p. 19. Geneva (1750)

Technology Adoption in the SME Sector for Promoting Agile Manufacturing Practices

Jawahar J. Rao and Vasantha Kumar

Abstract The Small and Medium Enterprises have always been constrained by technology and resources. In many instances, information technology has not penetrated enough to provide competitive advantage, and this is often due to skills and financial constraints. The study and analysis, in this paper is with regard to adoption of Information technology in the SME sector and the benefits obtained therein. The central entity is a large auto manufacturer, which has adopted Agile manufacturing practices and needs very tight integration of Information technology, manufacturing plans and market responses. This needs the downstream vendors, to respond in a timely manner, in order to achieve agility in the entire supply chain. This study involves, implementing information technology measure in select small enterprises and their response has been recorded in terms of lead time for production and inventory planning. This paper also discusses few more suggestions to achieve further improvement in the metrics.

1 Introduction

Manufacturing systems have influenced mankind for a very long time in terms of quality of life, comfort, and sophistication. This has provided a great deal of mechanization which has resulted in improved efficiencies. This has occurred due to rapid change in technology and innovation resulting in new products. The external environment, competition and market dynamics have resulted in rapid changes and introduction of new products and services.

In the past, economies of scale was the driving factor in manufacturing. The idea was to manufacture in large numbers and push it into the markets, in order to increase

J. J. Rao (✉) · V. Kumar
Department of Industrial Engineering and Management, Dayananda Sagar
College of Engineering, Bangalore, Karnataka, India
e-mail: jawahar_rao@yahoo.com

V. Kumar
e-mail: savkumar@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_69

revenues and use the plant capacity to the maximum. The monthly production targets were based on forecasts, which resulted in huge inventory of finished goods and work in progress. This has resulted in inflexibility in manufacturing and providing no possibility to reconfigure the systems, hence there was no scope to make changes based on market requirements or to make changes to product mix and model changes.

In the changing scenario, manufacturing companies need to adopt agile practices in order to compete with the best in the industry and also to be adaptive to change. This paper aims to study and analyze the challenges faced by the small and medium enterprises, in implementing the agile strategy and some of the levers that they need to apply, in order to succeed.

1.1 Literature Review

Research in agile manufacturing has occurred for more than 50 years and many authors have contributed immensely in this area. Authors in [1] have suggested several areas of research in agile, including design, information systems, supply chain and manufacturing systems. Agile manufacturing has opened up a new chapter in research and has kindled interest in many researchers. The authors have further commented on the above classification and mentioned their usefulness in each industry domain and their impact on manufacturing systems. Their work further indicates the benefits of agile manufacturing to users and customers alike.

Author [2] in her research has brought out the issues related to lean and agile manufacturing. The work has provided a comparison between the characteristics of lean and mass production as well as a comparison between mass production and agile manufacturing. The need to employ flexible manufacturing systems has been emphasized by the author. There is a compulsive need at the enterprise level to devise a manufacturing strategy in order to introduce new products and goods in a fast changing business environment, which is driven by market dynamics and competition. This defines and organization's ability to survive and thrive in a fast changing business environment.

Authors in [3] have opined that agile manufacturing will improve the competitiveness and provide an edge to business entities, who are quick to adapt and implement new technologies. The implementation of customer processes such as their marketing, design, supply chain will lead to an integrated delivery system, thus facilitating the agile philosophy.

Authors in [4] have commented that agile manufacturing systems will allow business entities to respond quickly and provide cost effective solutions in an uncertain environment and variable product demand, while helping organizations to launch new products in the market. This is very much applicable to unplanned products which are launched to meet changing customer requirements.

The models are applied to study the hypothetical decision of whether to invest in a dedicated, agile or FMS for engine and transmission parts machining. These decision models are the beginning towards development of practical business case tools that will help the manufacturing and service industry to avail the advantages of agile manufacturing systems. Authors in [5] have initiated cloud agile in the information systems area which is a paradigm shift from the normal manufacturing practices. The cloud agile in manufacturing provides users, the industrial production system as a service. This helps in offering the users various functional services that are available on the cloud, such as process design, factories virtualization, business integration, and production management. The advantage of these services on cloud is that users will be able to access them with minimal knowledge of the cloud technology. They do not have to be experts in the field. This system is advantageous since it offers the benefits of technology and models, such as SOA, clod computing and business process management.

Authors of [6] have stated that agile manufacturing will not be profitable without a strong supply chain system though, it is a fundamental part of agile. The agile supply chain aims to target the actual demand in the market place, employing the pull methodology. In other words organizations produce and sell only what is needed in the market. This is the pull methodology in agile, as against the traditional manufacturing methods using long term forecasts. Thus the philosophy of the agile supply chain is to sell what the market demands and not the goods that have already been produced.

The real paradigm shift required is to create the environment where change is the norm and is a challenge to be overcome [7]. The above Literature was reviewed to identify the gaps in the current research and evaluate the impact of different factors, contributing to agility in the Small and Medium enterprises in India.

2 Influencing Factors

A study of a few small and medium enterprises was undertaken to assess the inhibiting factors and possible methods to streamline the process from end to end. Though there has been an earnest effort to bring about agility, the following factors were sometimes difficult to overcome.

Some of these enterprises were following the traditional model of producing for a three-monthly forecast, since they had committed resources for such an arrangement. Their shop floor configuration was suited for mass manufacturing in a continuous fashion. There was dependency on third party vendors, to feed their assembly lines.

Some of them were producing downstream assembly components and were not aware of changes instantly. Lack of information and communication between customer and other third-party Vendors thereby created dependencies in the supply chain. In some cases there was poor adoption and integration of Information Technology tools.

3 Initiation to Agile—A Study

This study was conducted for a business entity, who is a large Auto manufacturer, wanted some of the downstream vendors to be agile, responding to quick changes to the current market and also be ready for the future markets [8]. A study of Business and manufacturing processes was conducted through interviews, visits and inspection, as well as Brainstorming. The study, brought out some of the glaring shortcomings in agile implementation, but some of the factors stood out as showstoppers in making these enterprises, quick and responsive to change. The start was made with an hybrid model, where the enterprises would start with the traditional model and quickly make changes to the production volumes as information regarding the product mix and variants would flow downstream. The most important factors identified were:

Information and Communication Technology
Design and Technology considerations
Collaboration
Environmental factors
Lead time to supply
Inventories

The effort here has been to enable these small enterprises, with necessary technology aids and tools in order to improve their business agility and measure their performance in terms of lead time and speed to market. Also, other indirect gains in terms of Inventory and costs were also recorded.

4 IT Enablement

The Auto enterprise has implemented ERP, for their end to end integration of their manufacturing and business processes. These five vendors at the downstream have been provided with terminals at their end which helps them to plan for production with a lead time of 15 days as per forecast. Since the customer works in the Agile mode, any changes in production numbers based on market requirements is immediately conveyed to the downstream vendors, through the system which enables them to make dynamic changes with very less lead time. For design considerations, especially for new component development, the suppliers can look to cloud applications for these services, using the following cloud models. The SOA architecture was used to integrate the vendors' legacy systems, with the ERP and billing systems. This was further tied up with MS-Project Management tool, for planning and tracking purposes. The agile cloud was enabled for design and new component development, which involved design of jigs, fixtures and tools. The following cloud deployments were considered based on their requirements and usage.

Public clouds are managed by third-party vendors who provide services to many different clients and they run a virtual factory for those clients. However, the information regarding the other clients and tasks carried out in the same factories or on the same machines, will not be known to the end users. In the Public cloud the end user will basically deal with edit processes and tracking changes. The new advances in technology has made, things easier for the SME sector, where these activities could be outsourced and costs are no more prohibitive.

Business entities use private clouds when the requirement is high data protection and security [9]. They need very little editing of the services provided by cloud such as computational resources and production, definition of its own services, machines, or even factories.

Hybrid clouds are a combination of both public and private clouds. In Hybrid clouds every customer partially owns and shares another server or machine, although in a controlled manner [10]. Hybrid clouds may be the key to achieving an external supply in scale form and under demand, but these clouds add the complexity of determining how to allocate tasks and processes across these different environments. Organizations may feel attracted to the promise of a hybrid cloud, but this option, at least initially, will probably be reserved only for simple applications without conditions, which do not require any synchronization or which will not require highly specialized or expensive equipment.

5 Results and Discussion

The five small enterprises were part of this study, and the results measured at regular intervals of 3 months. The overall gains were measured after one year, and tabulated below. The lead time in the traditional model was measured vis-a vis the agile methodology.

The observations from Table 1 also indicate that the agility and lead time reductions depend on the complexity of the components and their dependencies. In low complexity components, significant reduction in lead times was observed. In the C5, two more vendors were involved upstream in the processing, however reduction in lead time was obtained by way of effective communication and collaboration. The tools so far used are Interviews and Brainstorming, since it has been a small sample size.

6 Conclusion

The application of Information Technology is a big step forward in implementing an efficient agile framework. The cloud Agile can be used for design considerations, where small enterprises may not have all the technology and skills available with them. However communication and collaboration are also very important factors. In order to achieve, efficient.agile implementation, all stakeholders have to be in perfect sync, like in an orchestra. The large enterprises will have to treat their vendors as

Table 1 Supplier lead time in traditional vs agile supply chain

Supplier	Component	Lead time in traditional model	Lead time —agile	Inventory TM (days)	Inventory -agile	Remarks
S1	C1	30	15	15	7	There is scope for further inventory reduction, once there is process maturity
S2	C2	45	25	25	10	Lead time and Inventory is high when there is external dependency
S3	C3	30	20	10	7	Low complexity components can give maximum gains and no risk
S4	C4	20	10	15	5	Low value, C class items, result in lowest inventory
S5	C5	60	20	15	8	Multiple vendors, lead time reduction obtained through Collaboration

partners in business and need to maintain complete transparency and sharing of data well in time. Technology adoption and collaboration are low hanging fruits, which can give small enterprises an immediate advantage in implementing agile. There is further scope to reduce manufacturing lead times and incoming inventory once the efficiencies improve along the supply chain. The aim is to extend this study for a larger sample size, involving other factors such as market factors and collaboration, and bring out the advantages in terms of costs and Revenue.

References

1. Gunasekaran, A., Yusuf, Y.Y.: Agile manufacturing: a taxonomy of strategic and technological imperatives. Int. J. Prod. Res. **40**(6), 1357–1385 (2002)
2. Elkins, D.A., Hyuang, N., Alden, J.M.: Agile manufacturing systems in the automotive industry. Int. J. Prod. Econ. **91**, 201–214 (2004)
3. Lei, L., Ren, S.-J., Liu, W.-H., Wang, W.: Supply chain management model based on coordination. In: IEEE International Conference on Systems, Man and Cybernetics, Tucson, AZ, pp. 1806–1810 (2001)
4. Yu, T., Shen, X.: Study on quality information system for flexible production. In: IEEE International Conference on Automation and Logistics, pp. 1005–1009 (2008)
5. Macia-Perez, F., Berna-Martinez, J.V., Marcos-Jorquera, D., Lorenzo-Fonseca, I., Ferrandiz-Colmeiro, A.: A new paradigm: cloud agile manufacturing. Int. J. Adv. Sci. Technol. **45**, 47–53 (2012)

6. LiBo, Z.: A study on push—pull mode of supply chain based on system dynamics. In: IEEE International Conference on Grey Systems and Intelligent Services, Nanjing, pp. 1375–1380 (2009)
7. Hou, Z., Hu, J.: Information system evaluation model study of manufacturing enterprise. In: International Conference on Communication Systems and Network Technologies, Rajkot, pp. 521–524 (2012)
8. Huailiang, L., Pengwai, Z.: Modeling of agile manufacturing information system. In: International Conference on Technology and Innovation, 2006, ITIC 2006, Hangzhou, IET Publisher, ISSN: 0537-9989 (2006)
9. Sanchez, L.M., Nagi, R.: A review of agile manufacturing systems. Int. Prod. Res. **39**(16), 3561–3600 (2001)
10. Qinghong Shen, H.: Information system integration model of manufacturing enterprises based on object process methodology. In: International Symposium on Advanced Control of Industrial Processes, Hangzhou, pp. 609–614 (2011)

Leaf Recognition and Classification Using Chebyshev Moments

K. Pankaja and V. Suma

Abstract Earth contains millions of plants; each of this plant leaf has its own unique features. On related to their unique features plants leaf is used in different sectors in day to day life. Hence, proper identification of each leaf exhibits good result. According to the survey 50% of plant leaf is used in medical sector making medication for respective disease treatment. So plant leaf recognition plays a significant role. Many researches are conducted on leaf identification using different technology. This paper put forth an automatic leaf image identification model using image processing techniques. The proposed paper has been presented on leaf identification model, by using several feature extraction schemes. Feature extraction technique is carried out based on the texture, color and shape of the leaf images. The proposed model considered thirty classes of Flavia dataset with a total of 270 leaf images. The application of four different schemes for feature extraction increases the accuracy of the system up to 96.29%.

1 Introduction

Plants are one of the most valuable natural resources available on the earth. There is huge number of plant spices available in day to day life. Plants are playing significance role in agriculture, medicine, and forestry [1]. In medical field a variety of plants are used for treatment of life-threatening disease. Along with that Ayurvedic herbal field used a million numbers of plants in making medication for number of disease. So it is important to recognize a peculiar plant, based on leaf, playing a significance role

K. Pankaja (✉)
Cambridge Institute of Technology, Computer Science and Engineering, VTU,
Bengaluru, India
e-mail: pankaja.osr@gmail.com

V. Suma
Dayanand Sagar College of Engineering, Information Science and Engineering, VTU,
Bengaluru, India
e-mail: sumavdsce@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_70

in medical sector. Here a set of several plant leaf image are collected and analysis of each leaf image is essential [2]. Apart from medical field leaf identification is important in industries, helps the botanists in plant research along with that now a day's food engineers are focused on leaf recognition [3].

Normally leaves are identified by considering the color, structure and flower of the particular plants. But there are several leaves of plants which are similar in their structure. In such cases it is important to distinguish leaf; a lot of inquisition is done on leaf icon identification model using different image enhancement techniques. The similar structured leaf images are classified and recognize based on feature of the particular leaf images.

The proposed model is different from the previous models as this work use the combination of HSV color moments, DWT and Chebyshev feature extraction algorithm. The standard Flavia dataset is considered for the evaluation of the leaf recognition model. So far researchers have not used this combination for identification of a leaf. HSV color moments are used for the extraction of the features, to separate the particular color features of the respective image. Similarly features of texture and shape of the leaf are extracted using DWT and Chebyshev moments. SVM, one of the best methods is been used as a classifier.

2 Related Work

Recently many image processing technologies are designed and developed for particular leaf identification. Mzoughil et al. [4] designed automatic leaf recognition of the particular tree species using different sections of the input leaf images. In referred paper author mainly concentrated on different sections of the leaf images. Petiole detection is a more complex process involved in leaf recognition. In this paper authors designed a petiole detection model by considering two different image processing techniques. In combining with author worked on Base and apex detection plus leaf parts localization operation is conducted. A comparison operation is made with respective data set based on respective leaf of specific tree for identification. The experiment is conducted on 3070 icons of plant Leaves scan image samples.

Chemburkar et al. [5] presents an automatic leaf classification model. Morphological feature are mainly considered in this paper during leaf identification. The proposed paper is mainly concentrating on detecting herbal species, Prewitt edge detection image processing to recognize the respective edge of the input images. Along with that vein structure of the leaf icon is also detected. Comparison is made on both image processing output using neural network classifiers. The output which as the higher similarities is considered and finally based on that respective leaf is identified.

Yahiaoui et al. [6] has worked on automatic leaf identification model by using multiple leaflets. Leaves of the particular species are separated into compound and simple leafs, based on their respected shapes. The referred paper focused mainly on the compound leaves of species. The part-based shape decomposition method is implied for shape identification along with that texture of multiple leaflets are identified using histogram algorithm of Local Edge Orientation. The behavior of the respective pixel value is estimated by using Hough histogram function. Fusion operation is conducted on multiple leaflets. The Plant Leaves image dataset is considered during experiment analysis; the dataset totally includes 595 compound leaf images of 16 different species of plant.

Ab Jabal et al. [7] proposed a leaf identification model based on techniques of feature extraction and classify the respective input image using various classifier techniques. Each leaf has its related unique characteristics so the paper mainly explains the different techniques for feature extraction used in leaf identification. This paper also provides full study on various classifier model used in leaf identification with accuracy.

Patil et al. [8] proposed a growth rate identification model depending based on the respective leaf features. In referred paper a pre-processing techniques for digital image are applied better for system accuracy. The main intention of the proposed model is to consider the real-time leaf image; pre-processing techniques are applied to enhance the quality and color of input leaf images. The leaf edge information is extracted by watershed algorithm, along with texture feature and color feature using histogram, discrete wavelet transformation and Fast Fourier Transformation methods are used for texture feature extraction. Finally SVM classifier is designed in order to take a respective decision based features extracted from the leaf image.

Vijayashree et al. [9] explains the work conducted on medical plants which are used in Indian herbal. The author mainly focused on quality identification and purity of plant leaves used in medicine preparation. The texture of individual leaves are estimated and compared with the knowledge base, finally output is classified with the help of Principle Component Analysis. A judgement is made that the respective image of the leaf belongs to which division of plant resources. Based on the different characteristics of leaves the proposed image processing model reduces the problem present in leaves recognition.

Bora et al. [10] has worked on comparative analysis between two color spaces and with respect to color Image segmentation performance of HSV is better than $L*A*B$.

Gupta et al. [11] has compared the performance of Discrete Wavelets like Haar and Daubechies to provide a good reference to for application developers to choose a good compression wavelet.

The survey explores that the most of the work is done on designing leaf identification model based on single parameter or characteristics of each leaf images. So it is significance to design an integrated model where leaf identification done by considering all the parameters of leaves such as color, structure and shape which increase the accuracy of model during image analysis of the respective input image.

3 Methodology

The proposed system architecture is depicted in Fig. 1. The system architecture is designed in two operational phase, i.e., training and testing phase. Training system is nothing but creating a knowledgebase by using set of standard leaf samples. The proposed system is trained by 270 leaf samples with 30 different classes. In training section color, shape and texture features of leaf samples are collected by advanced feature extraction techniques, i.e., HSV color movements, DWT and Chebyshev movements and roundness. SVM classifier is trained using extracted numerical leaf samples features.

In testing section a real leaf image is considered as input. The selected RGB leaf image is converted into a gray scale image to reduce the functional complexity. The feature extraction techniques used in training section are used in the testing phase to extract the same set of features. The collected features of query image are compared with the features which are already stored in the knowledge base. SVM is the most effective binary classifier, which creates the hyper-plane based on similarity between the features. The features which are similar to corresponding category are recognized efficiently. The detail working of each functional block of proposed system is briefly depicted in below section.

3.1 *Image Pre-processing*

The input image quality is enhanced by using respective image pre-processing techniques which are suitable for the current operation. There are several pre-processing methods available in digital image processing, in the proposed application, RGB to gray scale conversion method is used. The leaf image pre-processing converts the three dimensional RGB color image into a two dimensional gray scale image as it

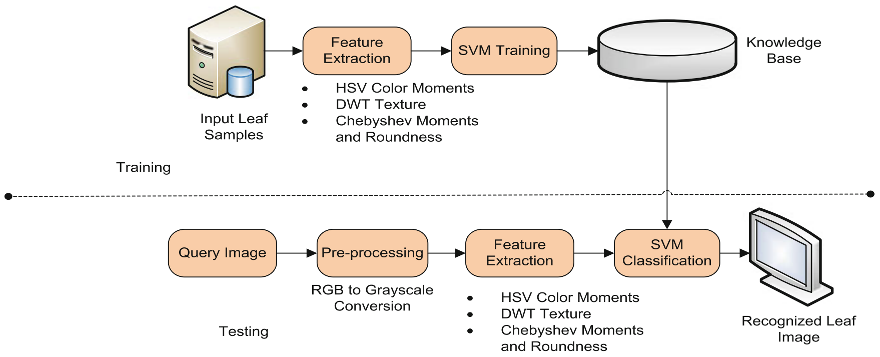


Fig. 1 Architecture of the proposed system

suite the further image analysis techniques. Equation (1) provides the mathematical equation involved in RGB image to gray scale converting

$$\text{Gray Image} = 0.2989 * \text{R} + 0.5870 * \text{G} + 0.1140 * \text{B} \tag{1}$$

3.2 Feature Extraction

Each leaf has its unique features and usually individual leaf is identified by its color, shape and texture, these different parameter of leaves are analyzed by using different image processing feature extraction algorithms. In proposed model the pre-processed image is subjected to feature extraction block. These various algorithms for feature extraction increase the accuracy and quality of the classifier. In designed model considered feature extraction methods are explained below.

3.2.1 Color Features Extraction

Color is generally an important feature of the leaf; each plant has its own distinct color combination. In image processing, each and every color is represented by respective mathematical model and with color models such as RGB and HSV. In RGB model, the image analysis is carried out based on three basic primary color of the input picture, i.e., Red, Green, and Blue. HSV which is defined as Hue, Saturation, and value and this proposed model mainly concentrates on HSV color model for features extraction [8]. Hue, Saturation and Value (Brightness) are considered as the important parameters during image analysis. Hue represents the combination of color component in form of angle, where the brightness of the images is described by the value (i.e. Brightness or Intensity). In combining with that, saturation explains the quantity of color mixed with the white. Bora et al. [10], discussed the mathematical equations from 1 to 10, for RGB to HSV color conversion model. Equations (2)–(10) provides the ranges between 0.255 of the respective color, i.e., RGB which is converted into 0 to 1 range by division operation as shown below

$$\text{R}' = \text{R}/255 \tag{2}$$

$$\text{G}' = \text{G}/255 \tag{3}$$

$$\text{B}' = \text{B}/255 \tag{4}$$

$$C_{\max} = \max\left(\text{R}', \text{G}', \text{B}'\right) \tag{5}$$

$$C_{\min} = \min\left(\text{R}', \text{G}', \text{B}'\right) \tag{6}$$

$$\Delta = C_{\max} - C_{\min} \tag{7}$$

HSV calculation is done by following equation

$$H = \cos^{-1}\left\{ \frac{\frac{1}{2}(\mathrm{R}-\mathrm{G}) + (\mathrm{R}-\mathrm{B})}{\sqrt{(\mathrm{R}-\mathrm{G})^2 + (\mathrm{R}-\mathrm{B})(\mathrm{G}-\mathrm{B})}} \right\} \tag{8}$$

$$S = 1 - \frac{3}{\mathrm{R}+\mathrm{G}+\mathrm{B}}[\min(\mathrm{R}, \mathrm{G}, \mathrm{B})] \tag{9}$$

$$V = \frac{1}{3}(\mathrm{R}+\mathrm{G}+\mathrm{B}) \tag{10}$$

The functional color feature extraction algorithm used in this paper is summarized in subsequent sections.

3.2.2 Texture Feature Extraction

Texture is one of the most significant features of the leaf and every leaf has its distinct texture making it difficult for the human being to analyze and identify the leaf derived from its texture. In this part of the research work, a Discrete Wavelet Transformation (DWT) algorithm is applied for input leaf image. Recently DWT has gained more importance in image processing as DWT function converts input image to the set of co-efficient values. These co-efficient patterns are defined as wavelets [11]. DWT dilations and shifting algorithm is applied to the single image prototype (also called as Mother Wavelet) to decompose it to form a useful and dynamic sub division of the input images or signals. These sub divisions are known as 'Daughter Wavelets'. Authors, of paper [12] has discussed the mathematical Eqs. (11) and (12) as below

$$h_{a,b}(t) = \frac{1}{\sqrt{a}} h\left(\frac{t-b}{a}\right) \tag{11}$$

For the equivalent input image X the wavelet transform is defined as

$$X_w(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} h * \left(\frac{t-b}{a}\right) x(t)\mathrm{d}t \tag{12}$$

The use of wavelet avoids the blackness present in the input signal and stores the sub-division in pixel blocks efficiently. The application of DWT decreases the calculation time which is basically timescale techniques that represent the input digital sample or pixel values in continuous form. In this proposed model, use of DWT helps to recognize the structure and texture of leaf image [12].

```
Algorithm :Color Feature Extraction
Input: 3D Color Image
Output: Color Features
    Step.1:  Assign a input image to variable imgIn
    Step.2:  Convert 3D into HSV plane
          i.e. outImg = rgb2hsv(imgIn);
    Step.3:  Find outImg size
    Step.4:  for i = 1: row
                 for j = 1: column
                      F1= round(outImg(i,j,1)*10+1;
                      F2 = round(outImg(i,j,2)*10+1;
                      F3= round(outImg(i,j,3)*10+1;
                      compute histogram of F1, F2 and F3;
                 end
          end
    Step.5:  Find out normalised histogram values
    Step.6:  Store normalised histogram value to output variable
End algorithm
```

3.2.3 Shape Feature Extraction

Image moment is one of the recent techniques of feature extraction which is used for analysis of image. Further, in image moments, the average weighted value of pixel intensity is considered based on that respective image shape or corresponding area is estimated. Though, there are different image moments methods, Chebyshev moments is one of the efficient discrete orthogonal image moments which reduces the numerical approximation problem and computational complexity problem present in Zernike's polynomial using Chebyshev moments. Additionally, Eq. (13) indicates that discrete orthogonal polynomials, i.e., $\{t_n(x)\}$ for given $N \times N$ 2D image must meet the following term as shown below

$$\sum_{x=0}^{N-1} t_m(x) t_n(x) = \rho(n, N)\delta_{mn}, \tag{13}$$

where $m, n = 0, 1, 2 \ldots N - 1$.

And t_n polynomial set and squared norm of t_n is denoted by $\rho(n, N)$. Equation (14) further shows that the orthogonal property of Chebyshev moment must meet the following given condition as,

$$\rho(n, N) = \frac{N\left(N^2 - 1\right)\left(N^2 - 2^2\right)\cdots\left(N^2 - n^2\right).}{2n + 1}, \tag{14}$$

where $n = 0, 1, \ldots N - 1$

However, the regularity of this above equation is represented further by Eq. (15) as given below

$$(n + 1)t_{n+1}(x) - (2n + 1)(2x - N + 1)t_n(x) + n\left(N^2 - n^2\right)t_{n-1}(x) = 0 \tag{15}$$

where $n = 1, \ldots N - 1$. When given image dimension is too large, i.e., N value than above represented Chebyshev polynomials patterns are unstable in their numerical form. This condition is easily identified when degree of t_n increase at N^n rate. This problem however is overcome by considering $t_n(x)$ Chebyshev polynomials with N^{-n} factor in image analysis. Resulting Chebyshev moments are hence represented by the Eqs. (16) and (17) as

$$T_{pq} = \frac{1}{\rho(n, N)\rho(n, N)} \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} t_p(x) t_q(y) f(x, y) \tag{16}$$

Here $pq = 0, 1, \ldots N - 1$.
Where,

$$\rho(p, N) = \frac{N\left(1 - \frac{1^2}{N^2}\right)\left(1 - \frac{2^2}{N^2}\right)\ldots\left(1 - \frac{p^2}{N^2}\right)}{2p + 1} \tag{17}$$

Also, authors of the paper [13] have used the Eqs. from (13) to (20) for computation of scaled $t_p(x)$ Chebyshev polynomials.

$$t_0(x) = 1 \tag{18}$$

$$t_1(x) = \frac{2x + 1 - N}{N} \tag{19}$$

$$t_p(x) = \frac{(2p-1)t_1(x)t_{p-1}(x) - (p-1)\left\{1 - \frac{(p-1)^2}{N^2}\right\}t_{p-1}(x)}{p} \text{ where } p > 1 \tag{20}$$

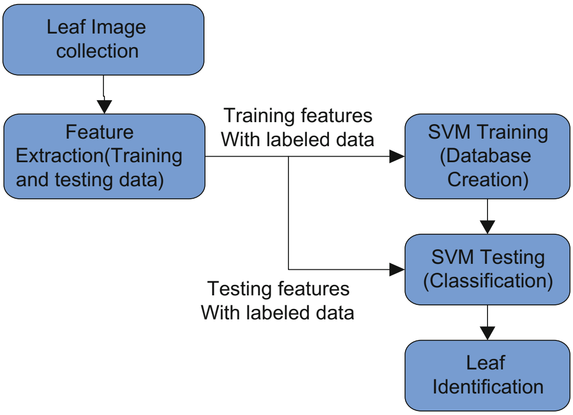
3.3 SVM Classifier

SVM classifier is mainly designed for regression and classification tasks, it is considered as the powerful binary classifier. The usage of classifier helps in detection of the leaf without any human interruption. SVM classifier create a hyper-plane which divide the feature of the respective input based on weight, hyper-plane is dot product of the set points with space constant. This clustering of supervised data compares testing data input to trained data, based on which decision is made. The operational flow of the respective SVM classifier is given in Fig. 2.

4 Results and Discussions

In this part of the research, the proposed method considers Flavia dataset with 30 species. 240 leaves are trained and 30 leaves used for testing. The color input image is

Fig. 2 Operational flow chart of SVM Classifier


Leaf Image collection
Feature Extraction(Training and testing data)
Training features With labeled data
SVM Training (Database Creation)
SVM Testing (Classification)
Testing features With labeled data
Leaf Identification

transformed to gray scale image in pre-processing model, which suit the further processing application of the proposed model. The leaf color combination is identified by using HSV model and the leaf texture is recognized by discrete wavelet transformation co-efficient. Finally, Chebyshev image moment algorithm is used for shape feature extraction. The intermediate proposed model result is represented in Fig. 3: (a) which represents input image, (b) represents gray scale image, (c) represents binary image, (d) represents color image and (e) represents classified result. Overall performance is evaluated using confusion matrix. This is a well-known matrix which is also known as error matrix. The matrix allows a visualization of concert of an algorithm. Every matrix row indicates the instances in a predicted class while every column indicates the instances in an actual class. Confusion matrix is a two row and two column tables that report the number of false positives, false negatives, true positives, and true negatives. Using confusion matrix this proposed model calculates the following parameters such as Sensitivity, Precision and True positive rate by using Eqs. (21)–(23).

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{21}$$

Sensitivity is the degree of awareness or responsiveness to external or internal changes or demands. Precision or positive predictive value (PPV) is given by,

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{22}$$

Precision is the mechanical or scientific exactness. Accuracy is given by,

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{23}$$

Accuracy is the amount of exactness of a quantity or expression.

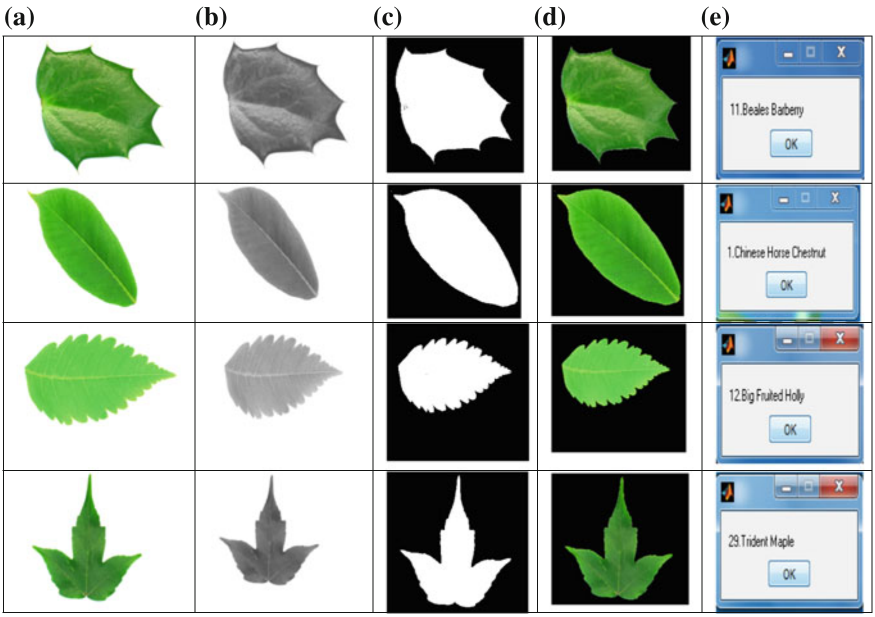


Fig. 3 **a** Input image **b** Gray scale image **c** Binary image **d** Color image **e** Classified result

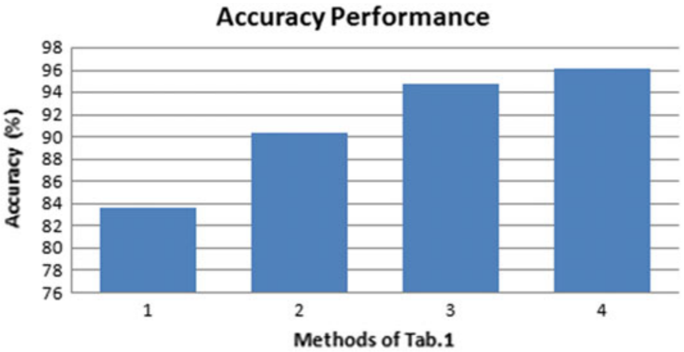


Fig. 4 Accuracy graph

The proposed model has evaluated totally 270 leaf samples, out of which true positive (TP) achieved is 210, false positive (FP) is 5, false negative (FN) obtained is 5 and true negative (TN) obtained is 50. Figure 4 and Table 1 depict the overall performance of the described system.

Table 1 Accuracy comparison of existing and proposed model

S. No.	Authors	Feature extraction algorithm	Classification algorithm	Overall accuracy (%)
1	Munisami et al. [3]	Color histogram	KNN classifier	83.5
2	Faizal et al. [7]	Probabilistic neural network & linear discriminant analysis	General Regression Neural Network (GRNN) classifier	90.312
3	Patil et al. [8]	DWT and FFT	SVM classifier	94.73
4	Proposed model	Chebyshev moments, HSV color moments, DWT texture	SVM classifier	96.29

5 Conclusion

Image processing has gained its significance in medical and health care fields. However, effective recognition and classification of medical plants is always a challenge. This paper therefore provides a computer-based automatic leaf identification model. The experiment is conducted on Flavia standard dataset where 30 different species are used in both testing and training phase. 270 leaves are considered, in which 240 leaves are trained and 30 samples are used for testing. This paper has used multiple feature extraction method that reduces the complexities, computational time and enhances the decision quality of classifier. The experimental result with classification accuracy 96.29% gives the working efficiency of the projected leaf identification model. Further, enhancements can be made by increasing the samples in the Flavia dataset and improve its robustness and accuracy. The work can be extended to other medicinal plant varieties. This work therefore paves the way towards enhancing the effectiveness to recognize, classify and further to use the medicinal plants for the betterment of the society.

References

1. Patil, A.A., Bhagat, K.S.: Plants identification by leaf shape recognition: a review. Int. J. Eng. Treads Technol. **35**(8) (2016)
2. Chothe, S.D., Ratnaparkhe, V.R.: Plant identification using leaf images. Int. J. Innovative Res. Sci. Eng. Technol. **4**(6) (2015)
3. Munisami, T., Ramsurn, M., Kishnah, S., Pudaruth, S.: Plant leaf recognition using shape features and color histogram with k-nearest neighbour classifier. In: Second International Symposium on Computer Vision and the Internet (2015)
4. Mzoughil, O., Yahiaoui, I., Boujemaa, N., Zagrouda, E.: Advanced tree species identification using multiple leaf parts image queries. In: 20th IEEE International Conference on Image Processing (2013)

5. Chemburkar, A., Sartape, A., Gawade, A., Somawanshi, P.: Automated tool for plant leaf classification using morphological features. Int. J. Eng. Comput. Sci. **3**(11) (2014)
6. Mzoughil, O., Yahiaoui, I., Boujemaa, N., Zagrouda, E.: Multiple leaflets-based identification approach for compound leaf species. In: International Workshop on Environmental Multimedia Retrieval (2014)
7. Ab Jabal, M.F., Hamid, S., Shuib, S., Ahmad, I.: Leaf feature extraction and recognition approaches to classify plant. J. Comput. Sci. **9**(10) (2013)
8. Patil, S., Soma, S., Nandyal, S.: Identification of growth rate of plant based on leaf features using digital image processing techniques. Int. J. Emerg. Technol. Adv. Eng. **3**(8) (2013)
9. Vijayashre, T., Gopal, A.: Authentication of leaf image using image processing technique. ARPN J. Eng. Appl. Sci. **10**(9) (2015)
10. Bora, D.J., Gupta, A.K., Khan, F.A.: Comparing the performance of *L*A*B* and HSV color spaces with respect to color images segmentation. Int. J. Emerg. Technol. Adv. Eng. **5**(2) (2015)
11. Gupta, D., Choubey, S.: Discrete wavelet transform for image processing. Int. J. Emerg. Technol. Adv. Eng. **4**(3) (2015)
12. Pandey, V.: Analysis of image compression using wavelets. Int. J. Comput. Appl. **103**(17) (2014)
13. Mukundan, R., Ong, S.H., Lee, P.A.: Discrete orthogonal moment features using Chebyshev polynomials. In: International Conference on Image and Vision Computing (2000)

Predictive Analytics Techniques Using Big Data for Healthcare Databases

Manjula Sanjay Koti and B. H. Alamma

Abstract Enormous amount of data is generated in the healthcare companies in the form of text, audio, video and electronic health record (EHR). These electronic health records are intricate and monotonous in nature and hence cannot be handled with traditional software or hardware nor with data management tools and methods. The application of big data in healthcare takes into account data explosion to obtain meaningful insights which helps in making effective informed decisions. Predictive Analytics tool helps in the proper diagnosis of the disease and appropriate treatments to be given to the patients who are suffering from certain diseases. The authors in this paper has discussed the usage of healthcare databases for big data analytics and the techniques used for analyzing the big data.

1 Introduction

Healthcare organizations generate huge amount of data from various sources such as clinical data, diagnosis data and doctor prescription which is highly unstructured and hence care is to be taken to structure the data [1]. There are lot of challenges which the healthcare industry faces. Healthcare domain generates data in the form of text, audio, video, Electronic Health Record (EHR), etc., which reveals the importance to develop the data analytics, and certainly this leads to shape up big data in healthcare. Further investigation of big data generates the predicted outcomes which can be used to better understand the trends in improving the healthcare, life expectancy and provide effective treatment at the preliminary stages at a very low cost. The significance to build up data analytics for future events are predicted through predictive analytics by means of data mining techniques. Decision support System makes use of data mining technique in turn to predict the diseases by utilizing medical data sets. Hence,

M. S. Koti (✉)
Sir M. Visvesvaraya Institute of Technology, Bangalore, India
e-mail: manjula.dsce@gmail.com

B. H. Alamma
Dayananda Sagar Academy of Technology & Management, Bangalore, India

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3_71

these medical data sets have gained popularity among researches worldwide [2]. Accordingly, the potential of predictive analytics in big data and the use of machine learning algorithm makes prediction task to be easier. Predictive Analytics tools help in the proper diagnosis and appropriate treatments to be given to the patients who are suffering from certain diseases [3]. Importance of predictive analysis lies in accurate diagnosis' of a particular disease, since doctors would have their own clinical decisions related to particular diseases. The authors have made a thorough investigation on the predictive analytics techniques.

This paper is organized as follows: Sect. 2 deals with importance of Predictive Analytics in healthcare, Sect. 3 discusses with literature survey, Sect. 4 presents usage of healthcare databases for big data analytics, Sect. 5 deals with the approaches used in dealing with big data, Sect. 6 deals with predictive analytics techniques and finally Sect. 7 presents the brief summary on the predictive analytics for big data on healthcare databases.

2 Importance of Predictive Analytics in Healthcare

Predictive analytics increases their accuracies of diagnosis' with appropriate use of predictive algorithms. For instance, whenever a patient visits the hospital with chest pain, it is hard to predict whether the patient needs to be hospitalized or not. In such situation, predictive algorithm would assist the doctors in accessing the likelihood of the patient status along with the aids of clinical judgement [4].

Predictive analytics will assist in preventive medicine and public health. In the early stages, most of the diseases can be prevented and hence Predictive analytics helps the physicians to recognize at-risk patients. This kind of information helps the patients to change their life style to avoid further risks. At the same time, decisions can be made by the physicians on the type of treatment that can be given for the patients [5].

Pharmaceutical organizations makes use of predictive analytics to meet the best needs of the public for medication. Industries will be given incentive to develop medication even for smaller groups. Drug which has been dropped since they were not been used by majority will be retrieved by the drug companies so that the patients utilizes this opportunity. Individuals will also receive treatments that will work for themselves instead of unnecessary medications. Application of predictive Analytics can also be extended to insurance and product cost which benefits both employers and hospitals. The employers can utilize the accurate algorithm which provides the benefits of healthcare for employees by storing the parameters of their workforce in order to obtain future medical cost.

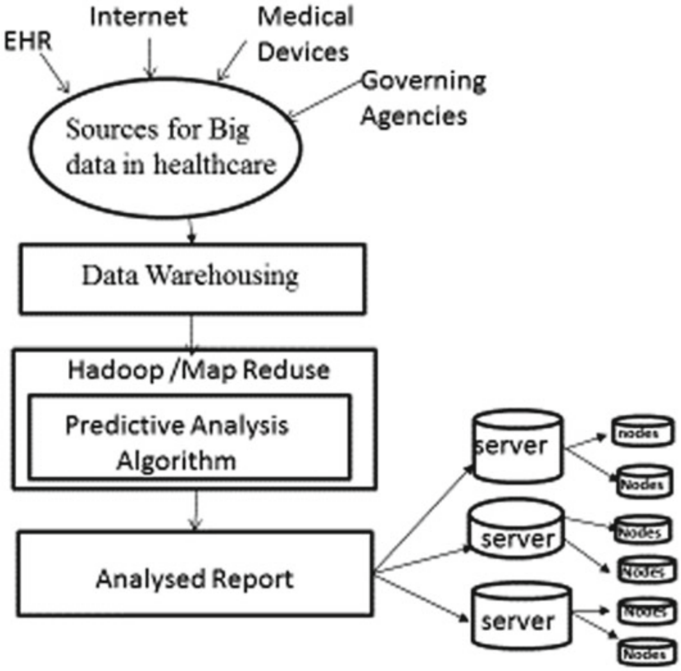


Fig. 1 Structural design of predictive analysis system

2.1 *Structural Design of Predictive Analysis System*

Importance of Predictive Analysis lies in predicting potential probabilities and trends. Figure 1 depicts the structural design of Predictive Analysis system. Basically, there are four important phases involved in designing Predictive Analysis system namely: sources for big data, datamining, Hadoop/MapReduce and report analysis [6].

Major phases in Predictive Analysis include:

Data Collection: It involves gathering the healthcare data which are highly unstructured and voluminous in nature from various resources such as electronic health record (EHR), Patient Health Record, clinical data, laboratories, insurance firms etc. and data exists in various formats and structure at various locations.

Data Warehousing: In this phase, the data collected from various sources is preprocessed to eliminate the redundancies and inconsistencies and thereby ensuring the quality of data [3].

Predictive Analysis: The providers of healthcare can expect and respond to the needs of the patients. This prediction helps in making financial and clinical decisions. Various predictive analysis algorithm can be used in Hadoop/MapReduce environment which in turn will predict the technical hitches bound with this and the various types of treatments that can be offered.

Analyzed Reports: After the careful examination of healthcare data via Hadoop/Map Reduce environment, the results are disseminated on multiple servers and this is reproduced at multiple nodes based on the geographical regions.

3 Literature Survey

In [1], the authors have discussed about the current developments in big data analytics with healthcare application domain, to improve the quality of life with regard to patients.

In [2], the authors have described about development of the decision system which helps doctors to take up the decision in various types of diseases. They have discussed analytical techniques to develop such systems.

In [3], the authors have explained the importance of structuring the data and its actual potential, since the data generated from healthcare industry is voluminous and highly unstructured. They have also discussed about the diagnosis of the chronic diseases which gives an insight about the big data from health care and this can be analyzed using predictive analytics methods by using tools like Hadoop/Map Reduce.

In [4], the authors have discussed about the complex classification type decision problems in healthcare and medicine fields. They have explained about the advanced analytical techniques by developing predictive models and using machine learning methods to diagnose a patient.

In [5], the authors have described about the adoption of electronic health record system by physicians and healthcare organizations. They have described the use of qualitative and quantity analysis to make decision using big data analytics.

In [6], the authors have discussed the importance of big data analytics in healthcare industries. They have described about the systematic approach for better outcomes of healthcare services to all population.

In [7], the authors have discussed big data analytics in healthcare, its benefits, framework and methodology. They have also discussed about its great potential in healthcare, by provide insights from very large volume data and obtain an efficient output by reducing cost.

4 Usage of Healthcare Databases for Big Data Analytics

4.1 Clinical Prediction Model

These days there is an exponential growth in healthcare data which makes us to get accustomed to crucial tools for exploring the data. Accordingly, these tools have a greater impact on the medical diagnostic system. Various Data Mining techniques Viz., association rules, decision trees and statistical methods such as Bayesian, probabilistic, logistic approach and regression models and machine learning techniques help in clinical decision measure. Further, the deployment of survival based techniques to measure the rate of survivability when a patient is suffering from a certain disease. But implementation of prediction models help the medical practitioners and customers in early diagnosis of diseases [5, 8].

4.2 Time Series Data Mining

Time series data is the temporal data which is measured with respect to time and data mining of such time series data is the process of discovery of hidden and explicit information through time series database. Further, temporal data mining of big data deals with automatic detection of hidden facts from data irrespective of dimensionality and complexity. Several temporal data mining pattern techniques such as temporal can be used to discover hidden facts and information from big data for future medical diagnosis [6].

4.3 Visual Analytics for Big Data

In visual analytics for big data, visualization of data will be made by integrating data from multiple resources. The patterns which are obtained through visual analytics helps in enhanced decision making with respect to clinical trials, effective marketing and increased revenue [9].

4.4 Information Retrieval

The main aim of information retrieval is to find the relative document in search process by querying, either structured or unstructured data using either web search or data search of user choice. The two main components of information retrieval includes indexing and retrieval of data, where indexing stores data in the form of index. It provides various opportunities for retrieving and querying big healthcare data for future prediction of diseases which in turn will be useful for doctors to effectively diagnose a particular disease [5].

5 Methodology

The various methodologies that can be utilized in handling the big data for healthcare databases include: [10]

Step1: The concept statement is developed by the healthcare team and then importance of the project is described. Further, the specialists from healthcare organization understands that there exists trade-off in cost, scalability, etc., after the approval of concept statement is the next stage.

Step2: In this phase, proposal development phase is initiated. Some of the important questionaries' such as the kind of problem being addressed, importance and interestingness of the problem, justification for using big data analytics approach, even though this approach is most expensive will be addressed.

Step3: In this step, the general project description statement is divided into a set of subtask, which is followed by the identification of dependent and independent variables. Data collection and transformation of data is done accordingly for the process of data analytics. At this juncture, one should take care of platform, selection/evaluation of tools. This is proceeded by application of big data analytics to data wherein these techniques can be utilized on huge data sets. Informed decisions are obtained through insights, where these insights are the results of big data analytics.

Step 4: In this phase, models which are constructed in previous stage, are implemented and their results are assessed and presented to the stakeholders for further action. One of the methodologies of big data is prediction analysis, which helps us to analyze the data and predict the result which helps us to better understand the ongoing trends of organizations [7].

6 Predictive Analytics Techniques

The patterns in historical and transactional data are analyzed by the predictive models to verify the risks associated along with opportunities. Predictive analytics analyzes voluminous data with different variables, with predictors being the core element measures an entity in order to predict its potential trends. Integration of multiple predictors leads to the development of predictive models that can be utilized to forecast future probabilities with certain reliability measures. Insights are produced which help the health care organization to understand the product cost and insurance and so on by combining the business knowledge and statistical analytical techniques [10, 11].

The main techniques for Predictive analytics include Data profiling and Transformations, Sequential Pattern Analysis and Time Series Tracking.

Data profiling and transformations involves the functions that alter the row and column values, investigate dependencies, formats of data, merge fields, aggregate records to form rows and columns. The relationship between the rows of data are identified through sequential pattern analysis. This method identifies the sequential items which occur frequently across all ordered transactions over time [12]. The sequence of values which are ordered at variable time intervals at a particular distance are termed as time series tracking.

The other common Predictive analytic techniques include Classification-Regression, Association analysis, Time series forecasting and so on [13].

Classification is one of the technique in data mining which samples the data into target class. It assigns the attributes of the data to the target class or the value of the numeric variable of interest is predicted [12]. The three phases involved in classification include training and testing by involving data with known class labels. Some portion of the data is utilized by the training data for construction of a model. Later, testing data will predict the class labels and accuracy of the model through which the model was developed through training data. In the deployment phase, the model is acceptable, then the model is deployed on unlabeled data. Popular classification algorithm are utilized for the development of classification models [4, 10].

Regression is a statistical technique, which model the association between one or more independent or dependent variable in which the values are continuous in nature. Multiple Linear regression is an advanced technique which uses multiple input variable for the deployment of more sophisticated models which are higher order polynomial equations [11].

Association analysis enumerates the co-occurrences of data items. This establishes relationships among diseases, health state and symptoms. Healthcare organizations use this technique to verify the fraudulent cases and abuse. The only evaluation factor considered in association technique is efficiency [8, 14].

Clustering tries to identify similarities between the data items. This technique forms the cluster from very large databases based on similarity measures and are mainly used in pattern recognition.

7 Conclusion

There is a shift in the healthcare industry from reporting facts to discovery of insights, leading to data-driven healthcare organizations. Accordingly, the healthcare value chain is significantly altered by big data from drug analysis to quality of patients care. Business intelligence data is used for forecasting and modeling. Selection of suitable data mining algorithms along with predictive modeling improves the search for healthcare data. The advanced technologies used by healthcare providers to gain insights from clinical datasets and make informed decisions has changed the Big Data Analytics. Big data analytics influences all the healthcare organization for better management by providing reliable services to people by understanding their needs. Data analysis can increase the process by using efficient tools to obtain the results, after analyzing the large data sets of healthcare along with their solutions.

References

1. Chauhan, R., Jangade, R.: A robust model for big healthcare data analytics. In: International Conference-Cloud System and Big Data Engineering IEEE@2016
2. Nilashi, M., Ibrahim, O., Ahmadi, H., Shahmoradi, L.: An analytical method for diseases prediction using machine learning techniques. Comput. Chem. Eng. **106**, 212 (2016)
3. Simpao, A.F., Ahumada, L.M., Galvez, J.A., Rehman, M.A.: A Review of Analytics and Clinical Information in Healthcare. Springer Science+Business Media, New York (2014)
4. Emanet, N., Oz, H.R., Bayram, N., Delen, D.: A comparative analysis of machine learning methods for classification type decision problems in healthcare. Decis. Analytics **1**, 6 (2014)
5. Simpao, A.F., Ahumada, L.M., Galvez, J.A.: A review of analytics and clinical informatics in health care. J. Med. Syst. **38**, 45 (2014)
6. Saravana Kumar, N.M., Eswari, T., Sampath, P., Lavanya, S.: Predictive methodology for diabetic data analysis in big data. Procedia Comput. Sci. **50**, 203–208 (2015)
7. Raghupathi, W., Raghupathi, V.: Big data analytics in healthcare: promise and potential. Health Inf. Sci. Syst. **2**(3), 2–10 (2014)
8. Koti, M.S., Srimani, P.K.: Medical diagnosis using ensemble classifiers-A novel machine learning approach. J. Adv. Comput. 9–11 (2013)
9. Thomas, J.J., Cook, K.A.: A visual analytics agenda. IEEE Comput. Graph. Appl. **26**(10–13), 38 (2006)
10. Zheng, J., Dagnino, A.: An initial study of predictive machine learning analytics on large volumes of historical data for power system applications. In: 2014 IEEE International Conference on Big Data, pp. 952–959 (2014)
11. Batra, A., Batra, U., Singh, V.: A review to predictive methodology to diagnose chronic kidney diseases. In: International Conference on Computing for Sustainable Global Development (INDA Com) (2016)
12. IBM: Data Driven Healthcare Organization Use Big Data Analytics for Big Gains (2013)
13. Geerdlink, B.: A reference architecture of big data solutions. IN: 8th International Conference for Internet Technology and Secured Transactions (ICITST) (2013)
14. IHT: Transforming Health Care Through Big Data Strategies for Leveraging Big Data in the Health Care Industry (2013)

Author Index

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 105,
https://doi.org/10.1007/978-981-13-1927-3