

Lakhmi C. Jain
George A. Tsihrintzis
Valentina E. Balas
Dilip Kumar Sharma *Editors*

Data Communication and Networks

Proceedings of GUCON 2019

Advances in Intelligent Systems and Computing

Volume 1049

Series Editor

Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences,
Warsaw, Poland

Advisory Editors

Nikhil R. Pal, Indian Statistical Institute, Kolkata, India

Rafael Bello Perez, Faculty of Mathematics, Physics and Computing,
Universidad Central de Las Villas, Santa Clara, Cuba

Emilio S. Corchado, University of Salamanca, Salamanca, Spain

Hani Hagras, School of Computer Science and Electronic Engineering, University
of Essex, Colchester, UK

László T. Kóczy, Department of Automation, Széchenyi István University,
Gyor, Hungary

Vladik Kreinovich, Department of Computer Science, University of Texas
at El Paso, El Paso, TX, USA

Chin-Teng Lin, Department of Electrical Engineering, National Chiao
Tung University, Hsinchu, Taiwan

Jie Lu, Faculty of Engineering and Information Technology,
University of Technology Sydney, Sydney, NSW, Australia

Patricia Melin, Graduate Program of Computer Science, Tijuana Institute
of Technology, Tijuana, Mexico

Nadia Nedjah, Department of Electronics Engineering, University of Rio de Janeiro,
Rio de Janeiro, Brazil

Ngoc Thanh Nguyen, Faculty of Computer Science and Management,
Wrocław University of Technology, Wrocław, Poland

Jun Wang, Department of Mechanical and Automation Engineering,
The Chinese University of Hong Kong, Shatin, Hong Kong

The series “Advances in Intelligent Systems and Computing” contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing such as: computational intelligence, soft computing including neural networks, fuzzy systems, evolutionary computing and the fusion of these paradigms, social intelligence, ambient intelligence, computational neuroscience, artificial life, virtual worlds and society, cognitive science and systems, Perception and Vision, DNA and immune based systems, self-organizing and adaptive systems, e-Learning and teaching, human-centered and human-centric computing, recommender systems, intelligent control, robotics and mechatronics including human-machine teaming, knowledge-based paradigms, learning paradigms, machine ethics, intelligent data analysis, knowledge management, intelligent agents, intelligent decision making and support, intelligent network security, trust management, interactive entertainment, Web intelligence and multimedia.

The publications within “Advances in Intelligent Systems and Computing” are primarily proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

**** Indexing: The books of this series are submitted to ISI Proceedings, EI-Compendex, DBLP, SCOPUS, Google Scholar and Springerlink ****

More information about this series at <http://www.springer.com/series/11156>

Lakhmi C. Jain · George A. Tsihrintzis ·
Valentina E. Balas · Dilip Kumar Sharma
Editors

Data Communication and Networks

Proceedings of GUCON 2019



Springer

Editors

Lakhmi C. Jain
Faculty of Engineering and IT
University of Technology Sydney
Broadway, NSW, Australia

Valentina E. Balas
Department of Automation
and Applied Informatics
“Aurel Vlaicu” University of Arad
Arad, Romania

George A. Tsirhrintzis
Department of Informatics
University of Piraeus
Piraeus, Greece

Dilip Kumar Sharma
Department of Computer Engineering
and Applications, Institute of Engineering
and Technology
GLA University
Mathura, India

ISSN 2194-5357

ISSN 2194-5365 (electronic)

Advances in Intelligent Systems and Computing

ISBN 978-981-15-0131-9

ISBN 978-981-15-0132-6 (eBook)

<https://doi.org/10.1007/978-981-15-0132-6>

© Springer Nature Singapore Pte Ltd. 2020

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

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, expressed 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

GUCON 2019 Organization

Chief Patrons

Mr. Suneel Galgotia, Chancellor, Galgotias University, India
Mr. Dhruv Galgotia, CEO, Galgotias University, India

Patron

Prof. Renu Luthra, Vice Chancellor, Galgotias University, India

General Chair

Mr. Deepak Mathur, Director-Elect, IEEE Region 10 (Asia and Pacific)

Conference Chair and Chairman, Oversight Committee

Prof. Rabindra Nath Shaw, Galgotias University, India

Conference Secretary

Prof. Priyabrata Adhikary, NHEC, India
Prof. D. Saravanan, Galgotias University, India

Technical Chairs

Prof. Yen-Wei Chen, Professor, Ritsumeikan University, Japan
Prof. Maria Virvou, HOD, Department of Informatics, University of Piraeus, Greece

Publication Chairs

Prof. George A. Tsihrintzis, University of Piraeus, Greece
Prof. Valentina E. Balas, University of Arad, Romania
Prof. Dilip Kumar Sharma, GLA University, Mathura

Honorary Chairs

Prof. Vincenzo Piuri, University of Milan, Italy
Prof. Georges Zissis, President, IEEE IAS
Prof. Lakhmi C. Jain, University of Technology, Sydney
Dr. Tamas Ruzsanyi, Ganz-Skoda Electric Ltd., Hungary

Honorary Co-Chairs

Prof. C. Boccaletti, Sapienza University, Italy
Prof. Mukhopadhyay, Ex-VC, Lingaya's University, India
Dr. Nishad Mendis, Det Norske Veritas, Australia
Dr. Akshay Kumar, Concordia University, Canada

Springer/GUCON Liaison

Dr. Aninda Bose, Senior Editor, Springer Nature

International Advisory Board

Prof. Valentina E. Balas, University of Arad, Romania
Prof. N. R. Pal, President, IEEE CIS

Prof. George A. Tsihrintzis, University of Piraeus, Greece
Prof. Yen-Wei Chen, Ritsumeikan University, Japan
Prof. Milan Simic, RMIT University, Australia
Prof. M. Paprzycki, Polish Academy of Sciences
Prof. Maria Virvou, University of Piraeus, Greece
Prof. Vincenzo Piuri, University of Milan, Italy
Prof. D. P. Kothari, Ex-Director, IIT Delhi, India
Prof. S. N. Singh, VC, MMMUT Gorakhpur, India
Prof. B. K. Panigrahi, Professor, IIT Delhi, India
Prof. R. K. Pandey, DG, NPTI, India

Technical Program Committee and Reviewers

Dr. A. R. Abhyankar
Dr. Aditi Sharan
Dr. Ajay Mittal
Dr. Sudhir Kumar Sharma
Dr. Ajai Jain
Dr. Alok Kushwaha
Dr. Amit Agarwal
Dr. Amalendu Patnaik
Dr. Anil K. Ahlawat
Dr. Anil K. Singh
Dr. Anuradha
Dr. Arun Kumar Verma
Dr. Aseem Chandel
Dr. Asheesh K. Singh
Dr. Ashutosh Dixit
Dr. Asif Ekbal
Dr. B. Dushmanta Kumar Patro
Dr. Baij Nath Kaushik
Dr. Bhaskar Biswas
Dr. Bharat Singh Rajpurohit
Dr. C. Patvardhan
Dr. C. Rama Krishna
Dr. C. K. Nagpal
Dr. Chandra Sekaran
Dr. Chiranjeev Kumar
Dr. Chittaranjan Hota
Dr. D. Bhagwan Das
Dr. D. A. Mehta
Dr. D. S. Kushwaha
Dr. D. S. Yadav

Dr. Desh Deepak Sharma
Dr. Dhram Singh
Dr. Dimple J. Gupta
Dr. Diwakar Bhardwaj
Dr. Girish Patnaik
Dr. Jai Govind Singh
Dr. Joy Deep Mitra
Dr. K. V. Arya
Dr. Kiran Kumar Pattanaik
Dr. Kishor K. Bhoyar
Dr. Komal Kumar Bhatia
Dr. Lalit Kumar Awasthi
Dr. M. K. Dutta
Dr. M. P. Singh
Dr. Madhavi Sinha
Dr. Manisha Sharma
Dr. Mohd. Rihan
Dr. Mayank Pandey
Dr. Munesh C. Trivedi
Dr. N. Badal, KNIT
Dr. Nanhay Singh
Dr. Narendra Kohli
Dr. Naresh Chauhan
Dr. Naveen Kumar
Dr. Neelam Duhan
Dr. Neeraj Tyagi
Dr. O. P. Verma
Dr. Pooja Jain
Dr. Pooja Pathak
Dr. Prabhat Ranjan
Dr. Prabhakar Tiwari
Dr. Prabin Panigrahi
Dr. Pragya Dwivedi
Dr. Pradeep Sharma
Dr. Pramod Kumar
Dr. Pramod Kumar Singh
Dr. Punam Bedi
Dr. R. K. Singh
Dr. R. S. Yadav
Dr. R. S. Rao
Dr. Rahul Rishi
Dr. Rajesh Prasad
Dr. Reena Dadhich
Dr. Ruchika Malhotra
Dr. S. P. Tripathi

Dr. Sapna Gambhir
Dr. Suneeta Agarwal
Dr. Sujoy Das
Dr. Sukomal Pal
Dr. Sunil Kumar Khatri
Dr. Tanveer Siddiqui
Dr. Tarun Shrimali
Dr. Vasudha Bhatnagar
Dr. Vishal Bhatnagar
Dr. Yashpal Singh
Prof. Herbert H. C. Lu
Dr. Senthilrajan Agni
Dr. Abhineet Anand
Dr. Anurag Baghel
Dr. Balamurugan Balusamy
Dr. Priti Bansal
Dr. Sonia Bansal
Dr. Annappa Basava
Dr. Rohit Beniwal
Dr. Vandana Bhasin
Dr. Rodrigo Bortoletto
Dr. John Moses Cyril
Dr. Pinaki Chakraborty
Dr. Sansar Chauhan
Dr. Rahul Chaurasiya
Dr. Surya Deo Choudhary
Dr. Anurag Dixit
Dr. Nripendra Narayan Das
Dr. Indrani Das
Dr. Aparna Datt
Dr. Parneeta Dhaliwal
Dr. Chandrakant Divate
Dr. Rajesh Dubey
Dr. Arman Faridi
Dr. Ankush Ghosh
Dr. Utkarsh Goel
Dr. Pallavi Goel
Dr. Amit Goel
Dr. Priyanka Goyal
Dr. Deepak Gupta
Dr. Suneet Gupta
Dr. Raza Haidri
Dr. Syed Shabih Hasan
Dr. Manas Hati
Dr. Brijesh Iyer

Dr. Manisha Jailia
Dr. Prashant Johri
Dr. Jegathesh Amalraj Joseph
Dr. Sandeep K. Singh
Dr. Vinay Kumar
Dr. Amita Kapoor
Dr. Sandhya Katiyar
Dr. Anvesha Katti
Dr. Ruqaiya Khanam
Dr. Aanchal Khatri
Dr. Shrawan Kumar
Dr. Devendra Kumar
Dr. Avneesh Kumar
Dr. Arun Kumar
Dr. Sanjeev Kumar
Dr. Vipin Kumar
Dr. Sanjay Kumar
Dr. Bhavnesh Kumar
Dr. Sandeep Kumar
Dr. Neetesh Kumar
Dr. M. Mohanraj
Dr. Ramakrishnan Malaichamy
Dr. Manas Kumar Mishra
Dr. Baibaswata Mohapatra
Dr. Thiagarajan Muthunatesan
Dr. Rashid Mahmood
Dr. Yogendra Meena
Dr. Gitanjali Mehta
Dr. A. K. Mishra
Dr. Keshav Nirjanjan
Dr. Manoj Panda
Dr. Sanjeev Pippal
Dr. V. A. Sankar Ponnapalli
Dr. Shiv Prakash
Dr. Sheetla Prasad
Dr. Mohammed Abdul Qadeer
Dr. R. Gunasundari Ranganathan
Dr. Ranjeet Kumar Ranjan
Dr. Rohit Raja
Dr. Bharti Rana
Dr. Mukesh Rawat
Dr. Navaid Zafar Rizvi
Dr. S. Pravindh Raja
Dr. Anil Kumar Sagar
Dr. Rajeev Sharma

Dr. Birendra Kumar Sharma
Dr. Shrddha Sagar
Dr. Jyoti Sahni
Dr. Mohd. Saifuzzaman
Dr. Kavita Saini
Dr. Kamalesh Sethuramalingam
Dr. Priestly Shan
Dr. Gavaskar Shanmugam
Dr. Dilip Kumar Sharma
Dr. R. P. Sharma
Dr. Mayank Sharma
Dr. Sudhir Sharma
Dr. Lokesh Kumar Sharma
Dr. Vishnusharma
Dr. Jitendra Singh
Dr. Girish Singh
Dr. Karan Singh
Dr. Harikesh Singh
Dr. Prashant Singh
Dr. Neetay Singh
Dr. Ajay Shanker Singh
Dr. Arun Solanki
Dr. Subhranil Som
Dr. Ritesh Srivastava
Dr. Vijayalakshmi Subramanian
Dr. Hardeo Kumar Thakur
Dr. Pradeep Tomar
Dr. Shashi Kant Verma
Dr. Sohan Kumar Yadav
Dr. Vinod Yadav
Dr. Dileep Yadav
Dr. Chandra Yadav
Dr. Emre Yay
Dr. Aasim Zafar
Dr. Usha Chauhan
Dr. Chetna Dabas
Dr. Sanjoy Das
Dr. Sumithra Gavaskar
Dr. Vimal Kumar

Preface

The book constitutes selected high-quality papers presented at the International Conference on Computing, Power, and Communication Technologies 2019 (GUCON 2019) organized by Galgotias University, India, in September 2019. It discusses the issues in electrical, computer, and electronics engineering and technologies. The selected papers are organized into three sections—cloud computing and computer networks; data mining and big data analysis; and machine learning and systems. In-depth discussions on various issues under these topics provide an interesting compilation for researchers, engineers, and students.

We are thankful to all the authors who have submitted papers for keeping the quality of the GUCON 2019 at high levels. We would like to acknowledge all the authors for their contributions and the reviewers. We have received an invaluable help from the members of the International Program Committee and the chairs responsible for different aspects of the workshop. We also appreciate the role of special sessions organizers. Thanks to all of them, we had been able to collect many papers on interesting topics, and during the conference, we had very interesting presentations and stimulating discussions.

Our special thanks go to Janusz Kacprzyk (Series Editor, Springer, Advances in Intelligent Systems and Computing Series) for the opportunity to organize this guest-edited volume.

We are grateful to Springer, especially to Dr. Thomas Ditzinger (Editorial Director, Applied Sciences and Engineering, Springer-Verlag), for the excellent collaboration, patience, and help during the evolution of this volume.

We hope that the volume will provide useful information to professors, researchers, and graduated students in the area of soft computing techniques and applications, and all will find this collection of papers inspiring, informative, and useful. We also hope to see you at a future GUCON event.

Broadway, Australia
Piraeus, Greece
Arad, Romania
Mathura, India

Lakhmi C. Jain
George A. Tsirhintzis
Valentina E. Balas
Dilip Kumar Sharma

Contents

An ECC with Probable Secure and Efficient Approach on Noncommutative Cryptography	1
Gautam Kumar and Hemraj Saini	
Probability Prediction Using Improved Method in Delay-Tolerant Network	13
Pradeep Yadav, Manuj Mishra and C. P. Bhargava	
Taxonomy of Cyberbullying Detection and Prediction Techniques in Online Social Networks	21
Madhura Vyawahare and Madhumita Chatterjee	
A Formal Modeling Approach for QOS in MQTT Protocol	39
E. Archana, Akshay Rajeev, Aby Kuruvela, Revathi Narayankutty and Jinesh M. Kannimoola	
Prediction of Gene Selection Features Using Improved Multi-objective Spotted Hyena Optimization Algorithm	59
S. Divya, Eranki L. N. Kiran, Madhu Sudana Rao and Pujitha Vemulapati	
A Compressive Family Based Efficient Trust Routing Protocol (C-FETRP) for Maximizing the Lifetime of WSN	69
Nandoori Srikanth and Muktyala Siva Ganga Prasad	
An Adaptive Genetic Co-relation Node Optimization Routing for Wireless Sensor Network	81
Nandoori Srikanth and Muktyala Siva Ganga Prasad	
A Novel Hybrid User Authentication Scheme Using Cognitive Ambiguous Illusion Images	107
Sumaiya Dabeer, Mahira Ahmad, Mohammad Sarosh Umar and Muneeb Hasan Khan	

Fault Classification in a Transmission Line Using Levenberg–Marquardt Algorithm Based Artificial Neural Network	119
Harkamaldeep Kaur and Manbir Kaur	
IoT Botnet: The Largest Threat to the IoT Network	137
Smita Dange and Madhumita Chatterjee	
Building a Trustworthy Ethical Approach to Cloud Computing	159
Ankita Sharma and Hema Banati	
Weighted Frequent Itemset Mining Using OWA on Uncertain Transactional Database	183
Samar Wazir, M. M. Sufyan Beg and Tanvir Ahmad	
Design of Customer Information Management System	195
Rohini Narayan and Gitanjali Mehta	
Modeling Machine Learning Agent for Interaction Conversational System Using Max Entropy Approach in Natural Language Processing	217
Anil Kumar Negi and Syed Imtiyaz Hassan	
Analysis of Energy Consumption in Dynamic Mobile Ad Hoc Networks	235
Indrani Das, Rabindra Nath Shaw and Sanjoy Das	
Improved ITCA Method to Mitigate Network-Layer Attack in MANET	245
Nilesh R. Marathe and Subhash K. Shinde	
Employing Machine Learning Models to Solve Uniform Random 3-SAT	255
Aditya Atkari, Nishant Dhargalkar and Hemali Angne	
A Design and an Implementation of Forecast Sentence Extractor	265
Benyatip Srichareon, Suparerk Manitpornsut and Prapas Pongdamrong	
Low Complexity Antenna Selection Scheme for Spatially Correlated Multiple Antenna Cognitive Radios	275
Sonali Chouhan and Tinamoni Taye	
Fair Comparative Analysis of Opportunistic Routing Protocols: An Empirical Study	285
Jay Gandhi and Zunnun Narmawala	
Distributed Optimal Power Allocation Using Game Theory in Underlay Cognitive Radios	295
Bhukya Venkatesh, Nadella Bala Sai Krishna and Sonali Chouhan	

Relay Selection-Based Physical-Layer Security Enhancement in Cooperative Wireless Network	305
Shamganth Kumarapandian and Martin James Sibley	
Analysis of Performance of FSO Link During the Months of Monsoon in Delhi, India	321
Sanmukh Kaur, Syed Zafar Ali Raza, Jaideep Khanna and Anuranjana	
Pectoral Muscle and Breast Density Segmentation Using Modified Region Growing and K-Means Clustering Algorithm	331
Jyoti Dabass	
Author Index	341

About the Editors

Dr. Lakhmi C. Jain, PhD, ME, BE(Hons), Fellow (Engineers Australia) is with the University of Technology Sydney, Australia, University of Canberra, Australia and Liverpool Hope University, UK.

Professor Jain founded the KES International for providing a professional community the opportunities for publications, knowledge exchange, cooperation and teaming. Involving around 5,000 researchers drawn from universities and companies world-wide, KES facilitates international cooperation and generate synergy in teaching and research. KES regularly provides networking opportunities for professional community through one of the largest conferences of its kind in the area of KES. www.kesinternational.org.

His interests focus on the artificial intelligence paradigms and their applications in complex systems, security, e-education, e-healthcare, unmanned air vehicles and intelligent agents.

George A. Tsirhrintzis is Full Professor and Head of the Department of Informatics in the University of Piraeus, Greece. He received the Diploma of Electrical Engineer from the National Technical University of Athens, Greece (with honors) and the M.Sc. and Ph.D. degrees in Electrical Engineering from Northeastern University, Boston, Massachusetts, USA. His current research interests include Pattern Recognition, Machine Learning, Decision Theory, and Statistical Signal Processing and their applications in Multimedia Interactive Services, User Modeling, Knowledge-based Software Systems, Human-Computer Interaction and Information Retrieval. He has authored or co-authored over 300 research publications in these areas, which include 5 monographs and 14 edited volumes. He is the Co-Founder and Co-Editor (along with Profs. Maria Virvou and Lakhmi C. Jain) of the Springer Series on Learning and Analytics in Intelligent Systems, the Editor-in-Chief of Intelligent Decision Technologies (IOS Press) and the International Journal of Computational Intelligence Studies (InderScience) and a member of the editorial boards of 8 additional journals.

Valentina E. Balas is currently Full Professor in the Department of Automatics and Applied Software at the Faculty of Engineering, “Aurel Vlaicu” University of Arad, Romania.

She holds a Ph.D. in Applied Electronics and Telecommunications from Polytechnic University of Timisoara. Dr. Balas is author of more than 300 research papers in refereed journals and International Conferences. Her research interests are in Intelligent Systems, Fuzzy Control, Soft Computing, Smart Sensors, Information Fusion, Modeling and Simulation.

She is the Editor-in Chief to International Journal of Advanced Intelligence Paradigms (IJAIP) and to International Journal of Computational Systems Engineering (IJCSysE), member in Editorial Board member of several national and international journals and is evaluator expert for national, international projects and PhD Thesis. Dr. Balas is the director of Intelligent Systems Research Centre in Aurel Vlaicu University of Arad and Director of the Department of International Relations, Programs and Projects in the same university.

She served as General Chair of the International Workshop Soft Computing and Applications (SOFA) in eight editions 2005-2018 held in Romania and Hungary.

Dr. Balas participated in many international conferences as Organizer, Honorary Chair, Session Chair and member in Steering, Advisory or International Program Committees.

She is a member of EUSFLAT, SIAM and a Senior Member IEEE, member in TC – Fuzzy Systems (IEEE CIS), member in TC - Emergent Technologies (IEEE CIS), member in TC – Soft Computing (IEEE SMCS).

Dr. Balas was past Vice-president (Awards) of IFSA International Fuzzy Systems Association Council (2013-2015) and is a Joint Secretary of the Governing Council of Forum for Interdisciplinary Mathematics (FIM), - A Multidisciplinary Academic Body, India.

She is also director of the Department of International Relations, Programs and Projects and head of the Intelligent Systems Research Centre in Aurel Vlaicu University of Arad, Romania.

Prof. Dilip Kumar Sharma is a Associate Dean (Academic Collaboration) and Professor at Department of Computer Engineering and Applications in Institute of Engineering and Technology, GLA University, Mathura, India. He is also Vice Chairman of IEEE Uttar Pradesh Section and IEEE Computer Society Chapter of IEEE Uttar Pradesh Section for the 2019. He has delivered/chaired more than 70 invited talks/guest lectures and chaired the technical sessions at various institutes/conferences. He has attended 32 short term courses/workshops/seminars organized by various esteemed origination and edited two books and worked as Guest Editor of International Journals of repute. He has organized more than 12 IEEE/CSI International/National Conferences and Workshops with the capacity of General Chair, Co-General Chair, Convener and co-convener etc. He has published more than 85 research papers in International Journals/Conferences of repute indexed in SCI, Scopus and DBLP databases and participated in 50 International/National conferences. He is Conferred 2018 Outstanding Section Volunteer Award by IEEE

Uttar Pradesh Section and also Significant Contribution Award by Computer Society of India in 2012, 2013, 2014 & 2017. His research interests are Web Information Retrieval, Data Analytics and Software Engineering. He has guided 2 PhD and 20 MTech Thesis, projects and various seminars undertaken by the students of undergraduate/postgraduate and he is currently guiding 5 PhDs and numbers of MTech/Thesis, Seminars and BTech Projects.

An ECC with Probable Secure and Efficient Approach on Noncommutative Cryptography



Gautam Kumar and Hemraj Saini

Abstract An Elliptic Curve Cryptography (ECC) is used on the Noncommutative Cryptographic (NCC) principles. The security and strengths of the manuscript are resilient on these two cryptographic assumptions. The claims on the Noncommutative cryptographic scheme on monomials generated elements is considered be based on hidden subgroup or subfield problems that strengthen this manuscript, where original assumptions are hidden and its equivalents semiring takes part in the computation process. In relation to the same, the research gap is well designed on Dihedral orders of 6 and 8, but our contributions are in security- and length-based attacks enhancement over Dihedral order 12, reported in work done. We modeled the said strategies and represent the ideal security concerns for applications.

Keywords ECC · Noncommutative cryptography · Monomials generations · Length based attacks

1 Introduction

In cryptography, the security algorithm and its measures are playing important role responsiveness, which has been considered as an integral part of computer science. Cryptography is a combined discipline of mathematics, computer science, electrical engineering, and physics. It is one of the foundations that give guaranteed secure communication in the presence of adversaries. Where, the strength and powerful computing techniques are most useful to avoid the threats and supports challenges. A lot of applications are available in the realistic sense for showing the essential requirements that contain to avoid adversaries to occur, the assurance of legitimacy,

G. Kumar (✉)
KL University, Hyderabad 500075, TS, India
e-mail: gautam21ujrb@gmail.com

H. Saini
Jaypee University of Information Technology, Solan 173234, HP, India
e-mail: hemraj1977@yahoo.co.in

protection of information from confession, protected message communication systems involved in transmission(s), and storage of information(s). The cryptographic algorithms are shown appropriateness in the full-fledged measurements with the proposed and/or available resources. But instead of the same, from a research point of view, the motivational issues on the algorithms with more impulsiveness and arbitrariness fondness are a guide for future research with an assortment on comparatively more and strong responses. In essence of cryptography, these are termed as private-key and public-key authentication and key exchange.

An immense revolution came through the use of Public-Key Cryptography (PKC), proposed by Diffie and Hellman [1]. The PKC's techniques introduced further in various forms, where on a variety of special features in Elliptic Curve Cryptography (ECC) [2, 3] is attracted the most attention in the area of cryptography. It is well available in the literature to show with marginal enhancement on the lower communication as well as computation costs. ECC provides better security and performance than RSA/DSA algorithms for equivalent security strengths on shorter key sizes [we followed National Institute of Standard and Technology (NIST) guidelines released in 2012]. Today, ECC is considered being tenable above the key length of 224 bits up to the year 2031, corresponding to the same 256 bits key lengths unsusceptible beyond 2031 and above key lengths are not defined but is secure, (Table 1). This table is indicating the RSA algorithm using 2048 and 3072 keyed sized bits for the same security strength for 224–255 and 256–383 varied lengths keyed and its an obvious relative performance advancements indicators.

From research points of view, all PKC's approaches are generalized on commutative-based principles, but some of the researchers were looking into the fact to generalize the cryptographic approach on noncommutative basis, and the given name for the same is noncommutative cryptography or non-Abelian cryptography. It is one of the approaches based on noncommutative nature, where it is mathematically based on random arithmetic operation star (*) (holds on rotation and/or reflection) on any of the noncommutative group G of $(G, *)$, where group G may be any Group elements, Ring elements, Semiring elements, or some algebraic structural elements or its combinations. According to its noncommutative naturalness properties for two elements or combinations of [if considered order be appropriate] a and b operations of G are not resembles the same results, such as $a * b \neq b * a$. It can be achieved on the combined principles from physics and mathematics that are producing the noncommutative natural generalization.

Table 1 Equivalent security for RSA versus ECC

RSA	ECC	Protection from attack
1024	160–223	Until 2010
2048	224–255	Until 2031
3072	256–383	Beyond 2031
7680	384–511	–
15360	512+	–

1.1 Related Work and Associated Issues

Noncommutative cryptographic approach keeps a solid backbone security enrichments and better performances than the existing approaches. Using noncommutative cryptography implementations in a number of applications are based on PKC's approaches such as on RSA/DSA, Diffie–Hellman, and ECC algorithms. For the cryptographic purposes, these are working efficiently in session-key establishment, en/decryption and/or in authentication systems on noncommutative too. The discrete logarithmic is acting as an intermediary strength near to non-negligible solutions. On behalf of the open opinion on security experts, a brief observation is presented here.

For solving a discrete logarithm problem (DLP) and integer factorization problem (IFP), Shor in 1994 [4] is given a competent algorithm on the quantum basis, so likely a representation of possible security breach on commutative-based cryptography. Further, Kitaev [5] considered the same as a special case on its DLP, and analyzed on its significance, called hidden subfield or subgroup problem (HSP). The general ideas from Paeng et al. [6], Joux and Nguyen [7], and Cocks [8] are one of the important steps in making the finite Abelian group's decision separations on cryptographic groups and its equivalents on quadratic residues. Magliveras et al. [9] designed PKC's using one-way functions and trapdoors infinite groups, therefore in 2002, Stinson observed sensibly on most of the PKCs that only belong on Abelian or commutative approach, whose forthcoming future intention may be susceptible in the arena. On behalf of same, Goldreich and Lee suggested don't put all the cryptographic generalizations in single "commutative group" only. Therefore, the reason was a clear indication to look at alternative cryptography for specific purposes; this was the opening of noncommutative cryptography. Noncommutative cryptography is a generalization of a commutative approach in such a way that it doesn't follow the commutative case properties, but those are analogous to be the commutative cases. Afterward, there are session-key establishment, en/decryption, and authentication schemes on noncommutative are generalized on a variety of schemes [10, 11]. HSP over elliptic curve cryptography-DLP (ECC-DLP) is comprehensively resolved by Proos and Zalka [12]. Lee [13] in 2004 organized quantum algorithms well on the random HSP for Noncommutative group elements and it was reporting well, with respect to braid group based attacks [14]. Further, Rotteler [15] suggested to use HSP over noncommutative with proven evidence are much harder and better in the adversaries presence. Cao et al. [16] used polynomials functions to build cryptographic scheme over noncommutative semirings or ring elements. Further, the protocol application was based on non-Abelian given by Kubo [17] on Dihedral order 6, which has been considered the initial order for this group and its construction is based on revolutions of three-dimensional approaches. Reddy et al. [18] build signature schemes over modular method on noncommutative groups and semirings. Moldovyan and Moldovyan [19] constructed the cryptographic implementations on four dimensions; the major intention was to generalize the security enhancement. Myasnikov and Ushakov [20] have the crypt analyzed on encrypted texts and the authentication schemes on the

hardness tests of the Conjugacy search problem on monoids elements. An algorithm is devised to solve the same problems and got anxious on the strategies. Svozil [21] recognized the metaphorical structures with hidden variable indecisiveness on non-contextual elements that can't be figured out on cryptanalysis, and it doesn't have any assembled proofs. Kumar and Saini [22] have shown the cryptographic applications on extra special group (ESG) that provides more robustness and unpredictable behavior as compared to all the known schemes using Noncommutative Cryptography on the extra special group (ESG) and applied the same in cryptographic schemes generations. Where the center of ESG is cyclic and its quotient belongs to nontrivial, i.e., the resultants are not equal to zero or identity to its group elements. Transitions from group elements to its equivalent semiring elements finish finitely on monomials (the proposed assumptions for a group and semiring elements are unique and irreversible on the proposed group) and contain all the algorithmic properties. It is designed the authentication and integrity schemes on Mono-morphism for a group and semiring elements on Dihedral Order 8 of ESG. Also, ESG defaults contain the authentication and integrity schemes on Heisenberg and Quaternion groups and finally it is illustrated on the exponential growth on Length Based Attacks and predicted almost to be unpredictable.

1.2 Motivation and Our Contribution

The issue related to security enhancement is one of the most motivational concerns, where monomials with semiring structures and Dihedral orders are presented on potential advantages to keep away the assumptions from various attacks. The monomials structured foundation is, in general, uses the equivalent semiring elements consideration takes part in the computation process, whereas main group parameters work in hidden, and it is based on polynomial modular reductions.

Our contribution highlights the monomials generations on the three dihedral orders of 6, 8, and 12. The Dihedral 6 is already presented so we didn't take into considerations, and Dihedral order 8 monomials generations for key-exchange, encryption-decryption, and authentication schemes presented in [22]. The virtual consideration for Dihedral order 12 is considered in the manuscript. In last, we have presented a scenario for length based attacks in order to investigate into Monomials Cryptographic generation approach.

1.3 Manuscript Organization

The manuscript is organized into subsequent sections, in the next section, it is presented with cryptographic assumptions on modular polynomials and further its hypothesis is presented on group and ring elements, in brief. In Sect. 3, preliminary knowledge of dihedral order 6 and 8 are presented from mathematical points of view

in justification of proposed strategies, which releases the significant contribution our proposed cryptography. In Sect. 4, a length-based attacks scenario is presented on input sequence generation and further presents scenario on attacks by the adversaries in reverse to find the original key. This represents security strength guarantees on enlarged search species.

2 Preliminaries

2.1 Noncommutative Assumptions on \mathbb{Z} Modular Strategies

A PKCs over the Noncommutative cryptography on polynomials with the semiring R elements is proposed by Cao et al. [16], and this scheme is generalized with the name of \mathbb{Z} -modular approach. The notation for a \mathbb{Z} -modular structure on ring r is $\mathbb{Z}(r)$, and its structural applications available on $\mathbb{Z}^+[r]$ for positive elements on noncommutative R and it is as well as applicable on negative $\mathbb{Z}^-[r]$, where $r \in R$ is not certain on general and monomials, where group and semiring are comprehensively applicable on \mathbb{Z} -modular.

2.2 The Basis of Noncommutative Cryptographic Algorithm

The concerns on security strengths are based on the following two assumptions:

- (i) **Conjugacy Decisional Problem (CDP):** The definition of CDP says on given two group elements a and b of group G , using the random secret chosen x to generate the other group elements that satisfies for $b = a^x$ or to generate the Conjugacy multiplicative inverse of: $b = x^{-1}ax$. It works in the forward direction.
- (ii) **Conjugacy Search Problem (CSP):** For a group G of elements a and b , that try to finds a secret x if there exists x in G such that $b = a^x$ or $b = x^{-1}ax$. It is a reverse process to determine the random secret key as x .

CSP is considered to be a one-way hash function generation, i.e., the designed algorithm(s) are not able to determine the other group elements values such as $a \rightarrow b^x$. In modern cryptography on Noncommutative, generalized assumption is completed enough to frustrate the cryptographers. Also, CSP is well known for its unrealistic nature to solve the same probably on polynomial time.

2.3 Monomials Used in \mathbb{Z} Modular Method

The \mathbb{Z} -Modular method is constrained to be monomials on the chosen polynomials on secrets parameters, i.e., original information of group elements are hidden with its equivalents ring/semiring elements on polynomials functions. Such participation Conjugacy assumptions are viewed as a special case. Conjugacy Search Problem is proposed under these considerations.

3 The Preamble to Dihedral Orders

3.1 Dihedral Orders 6

The dihedral is a virtual concept works on a finite set of group elements. After defined operations on it, the group elements show some specific variations that make the unique nature for cryptographic uses. The first initiated step for the noncommutative or non-Abelian group is dihedral order 6, denoted by D_3 , given by Uno and Kano in [23]. In which, three colored blocks such as Red, Green, and Blue is considered as an assumption, where three actions applies as “a: swap the first block and second block from left to right, b: swap the second and third block from left to right, e: leave the block as they are, and if two actions then do the operation from right to left as specified”. The set of operations works is as follows:

e: $RGB \rightarrow RGB$ or $()$, a: $RGB \rightarrow GRB$, b: $RGB \rightarrow RBG$, ab: $RGB \rightarrow BRG$, ba: $RGB \rightarrow GBR$, aba: $RGB \rightarrow BGR$

Here, the block operations are represented in the form of mathematics, with considerations on $R = 1$; $G = 2$; $B = 3$ and arranged the same in various group elements. Further, equivalent ring elements are assigned to group elements. It shows the center of group element results on a cyclic rotation having its quotient belongs to nontrivial elements, where variables or terms that don't result on the identity or zero elements.

$$\begin{aligned} \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix} &\rightarrow \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix} \rightarrow \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \\ \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix} &\rightarrow \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \end{aligned}$$

The cryptographic schemes such as key-agreement, encryption-decryption, and its authentication have presented on general and monomials generations in [24], and interested authors can go in detail with this reference.

3.2 Dihedral Order 8

The initial order for Dihedral order 8, denoted by D_4 is available in [25] and the use of this order for the cryptographic purpose is presented in [22]. These consist of the operations on the cyclic subgroup generation by rotations and reflections. For virtualization point of view Dihedral order 8 represented with an F on a square of glass with the alphabetic letter “ F ”. In the same, some defined operations have been considered, such as e acts as an initial assumption likely be an identity element, a is used for a rotation by 90° and b is used for reflection. To make use of cryptographic aspects, square movement makes a difference on 0° , 90° , 180° , 270° [clockwise rotations], are taken into its considerations and reflections on the other hand, as shown in Fig. 1.

This virtual concept we apply for numeric consideration for the use of cryptographic purposes. Another way to represent the dihedral order 8 concepts is still possible. The schematic representation is based on the square glass on three operations e , a , b and its corresponding mixed operations, represented in Fig. 2.

Finally, consider these group elements in a group from G_1 to G_8 , like e , a , a^2 , a^3 , b , ba , ba^2 , ba^3 , which have been used in cryptography for its specific uses such as session-key generation, en/decryption as part of its resultant. A similar idea for the same has assumed for Dihedral 12 on group elements from G_1 to G_{12} , detailed decryption is not available here, but we have considered. Interested authors may refer from Kumar and Saini [22].

Fig. 1 Symmetries of Dihedral order-8

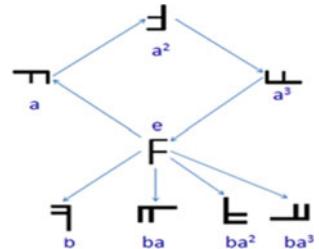
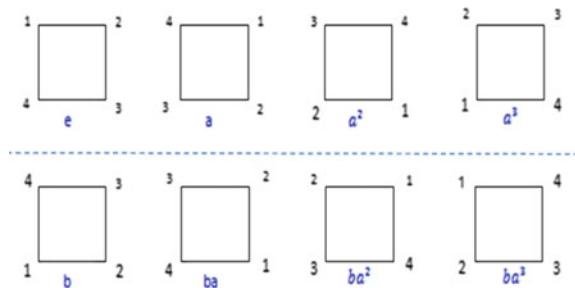


Fig. 2 Schematic representation on Dihedral 8



4 The Investigation into Length Based Attacks and Its Proposal

Length based attacks (LBA) is an approach to determine the user secret key; it works on word lengths, it is related to LBA, in Ruinskiy et al. [26], and in Myasnikov and Ushakov [27]. It is presented on Dihedral order 6 [24]. In these regards, it is one of the reverse procedures that try to recover the factors of conjugates. A good approach results in finding its Conjugator in the form of its group elements generation. The procedure is generating the Conjugators as follows for the Dihedral order of 6, 8, and 12, Fig. 3. On the input sequence and Dihedral order 6, there are 6 group elements and on the successful completion of this task a total of 36 elements to satisfy for the same. Similarly, for Dihedral 8, there are 8 group elements and a total of 64 elements. Finally, for the Dihedral 12 (hexagon element), there are 12 group elements and a total of 144 elements to satisfy for the same.

Our proposed approach is based on complexity enhancement for cryptographers (or complication) on Dihedral order of 6, 8, and 12. The group elements are $S_G = \{g_1^{\pm 1}, g_2^{\pm 1}, g_3^{\pm 1}\}$ for order 6, $S_G = \{g_1^{\pm 1}, g_2^{\pm 1}, g_3^{\pm 1}, g_4^{\pm 1}\}$ for order 8 and $S_G = \{g_1^{\pm 1}, g_2^{\pm 1}, g_3^{\pm 1}, g_4^{\pm 1}, g_5^{\pm 1}, g_6^{\pm 1}\}$ for order 12. The generation of input sequence on input $y = g_1 g_2^{-1} g_3 g_4^{-1}$, for length $n = 4$ for all the three orders as follows. On the assumption of any sequence of chosen input(s), perform operations on likely be on the $2k$ -ary tree. Where it does starts with an initial assumption word e , and generation of any further word/group elements depends on successful proceeding is one of the probable of its child generalized nodes. The successful accomplishment is based on chosen input y_n to length $y = y_1 y_2 \dots y_n$ traces likely as presented in Fig. 3. The n th-level contains elements on $(2k)^n$ leaf nodes. The leaf node of each one group is a potential element in any of y . The proposed work is difficult in finding its traces back on its cryptanalysis and/or decomposition of an encrypted message. The supporting group provides an unpredictable and robustness behavior on center and resultants in midair is/are rotates cyclic. Further, the assumptions are unique, irreversible and appropriate in sustaining to algorithmic properties. Therefore, for the proposed orders, it is assumed secure in reference to brute-force search.

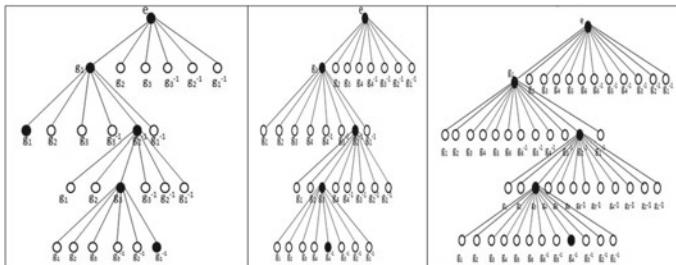


Fig. 3 The process of generating $y = g_1 g_2^{-1} g_3 g_4^{-1}$ on Dihedral 6, 8 and 12

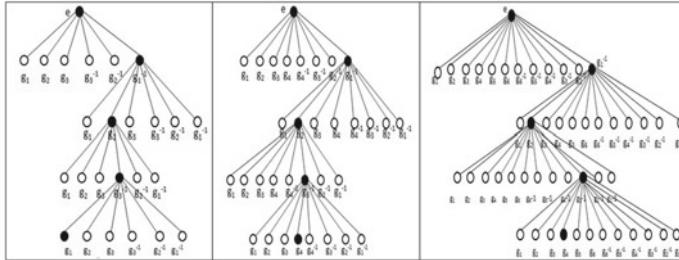


Fig. 4 Decomposition of length $y = g_1g_2^{-1}g_3g_4^{-1}$ on Dihedral 6, 8 and 12

Especially, when an attacker(s) tries to determine with equal child nodes, such as P and Q in same length, then the procedure have been created in such a fashion to fall for the same with insignificant solution. The general observation for Dihedral order 6, six candidates group elements forms at each level, so the time complexity work in the form to attack on the proposed strategy is $O(6^{2n})$ for all n word length, on the success or failure attempts. Next on average, 8 candidates (in Dihedral order 8) elements in each level for one group element, so the time complexity of this strategy is $O(8^{2n})$ for all n word length, on the success or failure attempts. Finally on average, 12 candidates (in Dihedral order 12, denoted by D_6) elements in each level for each group element, the time complexity of the attack algorithm is $O(12^{2n})$ for n length words, either on failure or success proceedings. The attack procedure is considered to be reversed searching the instance on the $2k$ -ary tree. In reference to shown Fig. 4, the decomposition on any lengths, the dark nodes are considered to be target nodes that forms paths, where this technique is suitable to find the path if it successful works. Therefore, we are able to enhance the robustness properties on its orders and accelerate the unpredictable behavior for cryptographic purposes. Its practical feasibility of the proposed idea is keeping a lot of benefits in noncommutative matrix operations on a finite group. Theoretically, the proposed approach is working; therefore interested authors and/or security agencies may apply this principle in various cryptographic applications, such as mobile techniques, online services, cloud in security, Internet of Things (IOT).

5 Conclusion and Future Scope

Due to tremendous demands on secured tools and techniques for various applications, our considered approach is one of the prime research concerns. The manuscript claims the Noncommutative cryptographic scheme on monomials generated elements. The monomials working principles are acting on Dihedral order of 6, 8, and 12. In regards to security and performance, these are reporting an immense contribution in the field of cryptography and making the proposal stronger based on the hidden subgroup

or subfields problem. For the adversary, the attacks like length based, cryptanalysis, and brute-force are likely being negligible to find.

As the proposed approach itself is a representation of polynomial functions that doesn't reveal secrets and/or finding polynomial for attacker is hard to find. The deployment considerations for applications are on high demand, designing for accelerating the algorithms, also in the area of security, is in tremendous demands.

References

1. W. Diffie, M.E. Hellman, New directions in cryptography. *IEEE Trans. Inf. Theory* **22**, 644–654 (1976). <https://doi.org/10.1109/TIT.1976.1055638>
2. V.S. Miller, Use of elliptic curves in cryptography. *Adv. Cryptol.* **218**, 417–426 (1986), dl.acm.org/citation.cfm?id=704566
3. N. Koblitz, Elliptic curve cryptosystems. *Math Comput.* **48**, 203–209 (1987). <https://doi.org/10.1090/S0025-5718-1987-0866109-5>
4. P.W. Shor, Algorithms for quantum computation: discrete logarithms and factorings, in *Proceedings of the 35th Annual Symposium on Foundations of Computer Science* (1994), pp. 124–134. <https://doi.org/10.1109/sfcs.1994.365700>
5. A. Kitaev, Quantum measurements and the Abelian stabilizer problem, in *Electronic Colloquium on Computational Complexity* (1996), <http://eccc.hpi-web.de/eccc-reports/1996/TR96-003/index.html>
6. S.H. Paeng, K.C. Ha, J.H. Kim, S. Chee, C. Park, New public key cryptosystem using finite non abelian groups. *Lect. Notes Comput. Sci.* **2139**, 470–485 (2001)
7. A. Joux, K. Nguyen, Separating decision Diffie-Hellman from Diffie-Hellman in cryptographic groups. *Cryptology ePrint Archive*, Report 2001/003 (2001), <http://eprint.iacr.org/>
8. C. Cocks, An identity based encryption scheme based on quadratic residues. *Lect. Notes Comput. Sci.* **2260**, 360–363 (2001)
9. S.S. Magliveras, D.R. Stinson, T.V. Trung, New approaches to designing public key cryptosystems using one-way functions and trapdoors in finite groups. *J. Cryptol.* **15**(4), 285–297 (2002)
10. K.H. Ko, D.H. Choi, M.S. Cho, J.W. Lee, New signature scheme using conjugacy problem. *IACR Cryptology ePrint Archive* 2002:168 (2002)
11. D. Grigoriev, I.V. Ponomarenko, On non-abelian homomorphic public-key cryptosystems. *J. Math. Sci.* (2002), cs.CR/0207079, [arXiv:cs/0207079](https://arxiv.org/abs/cs/0207079)
12. J. Proos, C. Zalka, Shor's discrete logarithm quantum algorithm for elliptic curve. *Quantum Inf. Comput.* **3**, 317–344 (2003), [http://dl.acm.org/citation.cfm?id=2011531](https://dl.acm.org/citation.cfm?id=2011531)
13. E. Lee, Braid groups in cryptology. *ICICE Trans. Fundam.* **E87-A**(5), 986–992 (2004)
14. D. Grigoriev, I. Ponomarenko, Constructions in public-key cryptography over matrix groups (2005), CoRR, abs/math/0506180, [arXiv:math/0506180](https://arxiv.org/abs/math/0506180)
15. M. Rötteler, Quantum algorithm: a survey of some recent results. *Inf. Forensic Entw.* **21**, 3–20 (2006), <http://link.springer.com/content/pdf/10.1007%2Fs00450-006-0008-7.pdf>
16. Z. Cao, X. Dong, L. Wang, New public key cryptosystems using polynomials over noncommutative rings. *Int. J. Cryptol. Res.* **9**, 1–35 (2007), <https://eprint.iacr.org/2007/009.pdf>
17. J. Kubo, The dihedral group as a family group, in *Quantum Field Theory and Beyond*, ed. by W. Zimmermann, E. Seiler, K. Sibold (World Science Publication, Hackensack, NJ, 2008), pp. 46–63, <http://www.worldscientific.com/worldscibooks/10.1142/6963>
18. P.V. Reddy, G.S.G.N. Anjaneyulu, D.V.R. Reddy, M. Padmavathamma, New digital signature scheme using polynomials over noncommutative groups. *Int. J. Comput. Sci. Netw. Secur.* **8**, 245–250 (2008), http://paper.ijcsns.org/07_book/200801/20080135.pdf

19. D.N. Moldovyan, N.A. Moldovyan, A new hard problem over noncommutative finite groups for cryptographic protocols, in *Lecture Notes in Computer Science*, vol. 6258 (Springer, Heidelberg, New York, 2010), pp. 183–194
20. A.D. Myasnikov, A. Ushakov, Cryptanalysis of matrix conjugation schemes. *J. Math. Cryptol.* **8**, 95–114 (2014). <https://doi.org/10.1515/jmc-2012-0033>
21. K. Svozil, Non-contextual chocolate balls versus value indefinite quantum cryptography. *Theoret. Comput. Sci.* **560**, 82–90 (2014)
22. G. Kumar, H. Saini, Novel noncommutative cryptography scheme using extra special group. *Secur. Commun. Netw.* **2017**, 1–21 (2017)
23. M. Uno, M. Kano, Visual cryptography schemes with dihedral group access structure, in *Proceedings of ISPEC'07* (Springer, 2007), pp. 344–359, <http://dl.acm.org/citation.cfm?id=1759542>
24. Z. Cao, New Directions of modern cryptography, in *Noncommutative Cryptography* (CRC Press, 2013)
25. B.C. Hall, *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction* (Springer, New York, 2003), <http://link.springer.com/book/10.1007%2F978-0-387-21554-9>
26. D. Ruinskiy, A. Shamir, B. Tsaban, Length-based cryptanalysis: the case of Thompson's group. *J. Math. Cryptol.* **1**, 359–372 (2007). <https://doi.org/10.1515/jmc.2007.018>
27. A.D. Myasnikov, A. Ushakov, Length based attack and braid groups: cryptanalysis of Anshel-Anshel-Goldfeld key exchange protocol, in *Lecture Notes in Computer Science*, vol. 4450 (Springer, Heidelberg, 2007), http://link.springer.com/chapter/10.1007%2F978-3-540-71677-8_6

Probability Prediction Using Improved Method in Delay-Tolerant Network



Pradeep Yadav, Manuj Mishra and C. P. Bhargava

Abstract Delay-Tolerant Network is a widely used network these days. Nodes are connected with each other and pass information between each other. The node that obtains the message accumulates it and advances its duplicate to another node it experiences. There are many parameters used to evaluate the Delivery Probability (DP) and storage of each node is also used because in DTN all the nodes should be capable of storing the message if the path is not available. In this paper, an improved method (EEA-PRoPHET) is proposed for increasing the probability prediction and transitivity of the nodes.

Keywords Delivery probability · DTN · EEA-PRoPHET

1 Introduction

1.1 Introduction

DTN is otherwise called astute systems. The DTNs are reasonable for work in the framework-less condition. This communication also allows the B/W wireless nodes in an odd environment. DTN or Disruption-Tolerant Systems DTN is the systems, which can be categorized as a subclass of MANET, where continuous connectivity from sender and receiver is not present at one point of time. Number of terminologies in the literature, such as ICMANET, Opportunistic Networks, challenged networks, or extreme networks has referred for DTN [1]. Each gadget on the endless sub arranges that contain the Internet makes utilization of this convention for information

P. Yadav (✉) · M. Mishra · C. P. Bhargava
ITM GOI, Gwalior, India
e-mail: er.pradepyadav0610@gmail.com

M. Mishra
e-mail: manuj.mishra02@gmail.com

C. P. Bhargava
e-mail: bhargavachandra1981@gmail.com

exchanges from source to goal with the negligible conceivable deferral and high dependability. End-to-end information exchange is the fundamental rule on which TCP/IP depends on.

DTN is an approach to manage PC masterminds outline that intends to paths the specific troubles in heterogeneous frameworks that knowledge nonattendance of predictable framework accessibility. DTNs engage data trade when adaptable centers are simply irregularly related. Since the availability isn't required to be steady in DTN, it utilizes what is known as a store-convey and forward directing system. In this, the halfway versatile hubs convey information bundles when they get it and forward it to the following hub as and when contact is built up. As DTN relies upon versatile hubs to convey information, the execution of directing the information exclusively relies upon whether the hubs [2] interact with one another or not.

DTN is a networking administration engineering that is intended to give correspondences in heterogeneous situations, where the system would be visit and enduring interruptions and high bit error rates that could regularly debase the execution. Interruption-tolerant system is developed from MANET. It is inadequate and irregular associated organize where dependable correspondence end to end availability isn't accessible for message transmission. Sensor-based networks, Wireless networks (WN), Terrestrial wireless networks (TWN), and submerged acoustic systems with delays are the models of DTN. New layer presented named "Package layer" in the engineering of DTN, is over the vehicle layer and underneath of use layer utilized for store and forward of messages. This engineering has various difficulties like absence of framework, separation, disturbance, and absence of assets. In these "Store and Forward" system, every hub [3] in DTN stores approaching messages in the cradle and advances it with regards to the goal hub or closer of these coveted goals among hubs.

1.2 Routing Protocols

There are different kinds of routing protocols exist in DTN. Basically, there are two different classifications of directing data in DTN, flooding methodology and forwarding procedure. Here, we have to find out the best route to transmit data from source to destination.

- a. Single Hop Transmission protocol.
- b. Binary spray and wait.
- c. Location based routing strategy.
- d. PROPHET (Probability Routing Protocol using History of Encounters and Transitivity).

The technique for store and forward is exceptionally practically equivalent to the genuine postal administration. Each letter needs to go through an arrangement of post workplaces; here it is handled and sent, before achieving the goal. Here, the

aggregate message or a lump of it is moved and put away in hubs progressively until the point that it achieves the goal.

2 Literature Work

2.1 Literature Review

In this section, the work done by several researchers is being discussed.

Md. Sharif Hossen [4], In an ICMN scenario for every the duplication base DTN routing protocol in particular pandemic, Prophet, Max-Prop, Spray and Wait, and Spray and Focus, the implementation is assessed against various message age rates and number of portable rates respectably. Delivery probability, normal inertness, and overhead proportion are three execution analyzer networks. For these, one simulator is utilized as an instrument.

Cao [5] in his work asserted geographic directing plan for beating the test to acquire the continuous area of goal in DTNs. In this work, they at first figure the closeness to the advancement expand assessed for objective by considering the adaptability vector of versatile center. At that point, the memo is recreated by the calculations named as Reach Phase and Meet Phase individually, for quick message conveyance. In the event that the message is out of the development go then it utilize the Reach stage and on the off chance that it is inside the range then it utilize Meet stage. The recreated messages are again under organized broadcast with the assistance of planning plan among Reach and Meet Phase, for improving the steering execution.

Grasic [6] proposes the current DTN assessment hones through an intensive and writing study. Creators demonstrate some shortcoming utilized in the assessments and proposes a model for assessment of DTN steering plans that track the essential sources of info, which should be chivalrous in the assessment procedure. Creators initially outline the related work then sequential review and after that proposed DTN assessment demonstrate. The model took three classes of contributions to produce the outcomes in the reproduction situations.

Ferreira [7] in his work proposed a system administration answer for VDTN utilizing the standard SNMP tradition. An observing framework is extremely helpful to confirm the correct system working, checking conceivable system inconsistencies, and to gather measurement information for the system head. The SNMP application is inserted into VDTN terminal hub application.

Beuran et al. [8] in their work proposed an imitating tested expected for DTN application and convention tests. Examined in detail the principle changes that were essential with a specific end goal to make conceivable DTN investigates the QOMB, a nonexclusive reason remote system imitating tested are multi-interface support, fault injection mechanisms, etc.

3 Work

3.1 Proposed Work

In the proposed work the performance of an Energy-Aware PRoPHET is being improved. In the proposed work, the calculation of Delivery Probabilities (DP) is performed. Then energy and node's available buffer are taken into consideration. Here, the remaining energy is used to find the encounter rate. If the energy is less and DP is higher than there is a chance of node die before reaching the destination node. So, energy is additionally exceptionally fundamental part to enhance the general execution. There are many parameters used to evaluate the Delivery Probability (DP) and storage of each node is also used because in DTN, all the nodes should be capable of storing the message if the path is not available. When the node encounters another node, then it checks the remaining energy and free buffer. This method is used to ignore the issue of storing extra information about each node.

3.2 Proposed Algorithm

Let $T(a, b)$ be a DP node (a) has for node (b) and T_{Enc} be the encounter value. Then, a offset which is simplified in every counteract is intended as given below

$$T(a, b) = T(a, b)\text{old} + (1 - T(a, b)\text{old}) * T_{Enc} \quad (1)$$

P_{Enc} is calculated by using the following formula:

$$P_{Enc} = \begin{cases} Tm * \left(\frac{TLE(b)}{TE} \right) & \text{if } f0 \leq TLE(b) \leq TE \\ . & \end{cases} \quad (2)$$

where TLE represents the time for the last encounter with node b and TE is the expected interval of time for connections.

- Step: 1 In the work first of all, delivery probability would be calculated and then new neighbor node would be calculated after that apply decay values on DPs.
- Step: 2 Node (a) send information to encountered node (b) such as DP_a, S_{va}, E_a. Node (b) receive data and perform further evaluation.
- Step: 3 If E_a is less than E_b then calculate probability of node a and b and size of free buffer (B_f) f node b
- Step: 4 If (T_a < T_b and B_f > Sizemsg) add message in the send list else B_f = B_f—Sizemsg
- Step: 5 End message.

4 Performance Metrics

4.1 Delivery Probability

It is a ratio of overall number of message conveyed to the target and the no. of messages created.

$$\text{Delivery Prob} = \frac{\text{Number of messages conveyed}}{\text{Number of messages created}} \quad (3)$$

4.2 Overhead Ratio

It is a distinct as number of required repeated packets is spread for delivering over the one packet to the destination.

$$\text{Overhead Ratio} = \frac{L - D}{D} \quad (4)$$

where L represents quantity of information send by transfer node and D represents quantity of messages conveyed to the goals.

These are some different performance metrics which shows the measurement in terms of quantity as well as quality wise.

5 Results

5.1 Results in Tabular Form

Here, the result is shown in the tabular form. Delivery probability has been calculated using different numbers of nodes (Table 1 and Fig. 1).

Table 1 Table shows the comparisons between both the methods

No. of nodes	EA-PRoPHET	EEA-PRoPHET
100	0.4722	0.4743
200	0.5171	0.5206
300	0.4911	0.5034
400	0.4552	0.4586
500	0.4204	0.4272

Fig. 1 Delivery probability against different numbers of nodes

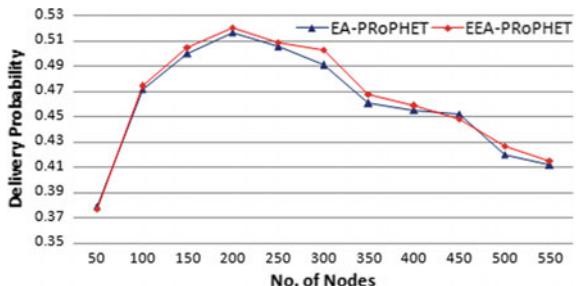
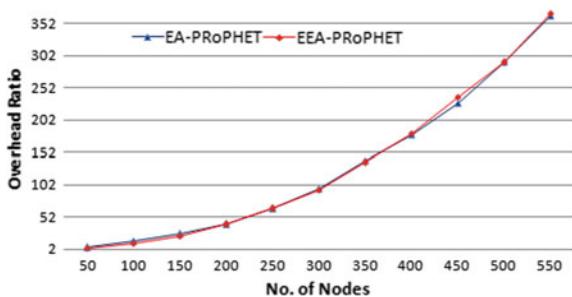


Table 2 Table shows the comparisons between both the methods

No. of nodes	EA-PRoPHET	EEA-PRoPHET
100	14.6372	11.6012
200	40.809	40.332
300	95.7483	93.8697
400	179.3579	180.5522
500	292.0537	291.2608

Fig. 2 Results with the help of pie chart



In this figure, comparison between both the methods is shown. By the pie chart, we can see that the proposed method shows better results.

Similary by using the performance metric of overhead ratio, the EEA-PRoPHET method shows the best result in comparison to EA-PRoPHET (Table 2).

Here, the results is shown with the help of pie chart (Fig. 2).

6 Conclusion and Future Work

DTN provides unique features of intermittent connectivity between nodes, which helps in communication when connection breaks and reconnect after some time and makes routing different from traditional network. The proposed work is better to perform the prediction delivery and provide better transitivity among the nodes.

References

1. A.S. Patil, P.J. Kulkarni, Exploiting social relations for efficient routing in delay tolerant network environment. *Int. J. Comput. Sci. Eng.* **6**(1) (2018). E-ISSN: 2347–2693
2. N. Dayanand, A. Vidhate, Improved routing protocol for delay tolerant network. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **6**(4) (2016). ISSN: 2277 128X
3. Y. Patel, T. Patel, A survey on DTN routing protocols. *Int. J. Eng. Dev. Res.* **3**(4) (2015). ISSN: 2321-9939
4. S. Krug, J. Seitz, Challenges of Applying DTN Routing Protocols in Realistic Disaster Scenarios. 978-1-4673-9991-3/16/\$31.00 ©2016 IEEE
5. Y. Cao, Z. Sun, N. Ahmad, H. Cruickshank, A mobility vector based routing algorithm for delay tolerant networks using history geographic information, in *2012 IEEE Wireless Communications and Networking Conference: Mobile and Wireless Networks*, pp. 2757–2762
6. S. Grasic, A. Lindgren, An analysis of evaluation practices for DTN routing protocols, in *CHANTS'12*, Istanbul, Turkey, pp. 57–63, 22 Aug 2012
7. B.F. Ferreira, J.N. Isento, J.A. Dias, J.P. Rodrigues, L. Zhou, An SNMP based solution for vehicular delay-tolerant network management, in *Global Communications Conference (GLOBECOM)* (2012)
8. R. Beuran, S. Miwa, Y. Shinoda, Network emulation testbed for DTN applications and protocols, in *INFOCOM* (2013)
9. Y. Cao, Z. Sun, N. Wang, F. Yao, H. Cruickshank, Converge-and-diverge: a geographic routing for delay/disruption-tolerant networks using a delegation replication approach. *IEEE Trans. Veh. Technol.* (2013)
10. F. De Rango, S. Amelio, P. Fazio, Enhancements of epidemic routing in delay tolerant networks from an energy perspective, in *2013 9th International Wireless Conference on Communications and Mobile Computing (IWCMC)*, pp. 731–735, July 2013
11. B.B. Bista, D.B. Rawat, A robust energy efficient epidemic routing protocol for delay tolerant networks, in *2015 IEEE International Conference on Data Science and Data Intensive Systems*, pp. 290–296

Taxonomy of Cyberbullying Detection and Prediction Techniques in Online Social Networks



Madhura Vyawahare and Madhumita Chatterjee

Abstract Online social networking sites have become very popular in this era due to easy accessibility of Internet. This popularity leads to continuous availability of multiple users, which resultantly attract more criminals and hence increasing insecurity in OSN. Different types of crimes are committed for multiple reasons in cyber realm by taking assistance of cyber technology. This insecure environment of OSN needs attention to prevent the damage caused by these crimes to society. Cyberbullying is reported as one of the harmful crimes causing psychological damage to victims. Cyberbullying has dangerous effects on the victim, which may also lead the victim to suicidal attempt. Victims of cyberbullying are usually afraid or embarrassed to reveal about their harassment. It has become a necessity to detect and prevent cyberbullying. Many researchers are working in multiple directions to achieve best results for automated cyberbullying detection. We have done a broad survey of all recent techniques proposed by researchers for cyberbullying detection and prediction. In the paper, we have presented taxonomy of multiple methods being used for cyberbullying detection. We also have presented a comparative analysis and classification of the work done in recent years.

Keywords Cybercrime · Cyberbullying · Online social network · Machine learning

1 Introduction

In this era of technology, the popularity of online social networks (OSN) is increasing not only among technologically strong people but also among nontechnical people. Availability of internet is one of the major reasons behind the high utilization of

M. Vyawahare (✉)

Pillai College of Engineering, Mumbai University, Navi Mumbai, India

e-mail: madhura.vyawahare@gmail.com

M. Chatterjee

Pillai HOC College of Engineering and Technology, Mumbai University, Rasayani, India

e-mail: mchatterjee@mes.ac.in

OSN. OSN have no generalized definition but it is defined by authors Boyd et al. [1] as a web service that allows an individual to do three things: (A) Generate a public or semipublic profile in a specific system, (B) Create a list of users to interact with and browse through the list of contacts and (C) See what was done by others within the system. People are fond of OSN for many purposes like entertainment, social contacts, fun, fame, advertisement, business. OSN has also become a major platform for cybercriminals to perform several types of crimes. Huge amount of data is present on social media which is being utilized for several criminal purposes. Rate, types, and complexity of attacks are increasing drastically due to such big platforms where people are available relentlessly [2].

Cyberbullying is one of the harmful cybercrimes, which is recently reported as a crime causing tremendous psychological damage to the victim. Cyberbullying is defined as “willful and repeated harm inflicted through the use of electronic devices” [3]. Detection of this crime is considered largely by many researchers. It is a relatively new area, but the research progress is exponentially incremental due to: (A) The increasing rate of occurrences of cyberbullying observed in OSN, (B) The damage it is causing to society. Detection of cyberbullying is considerably focused; at the same time, few researchers have also targeted prediction of cyberbullying by analyzing and relating previously available data on social networks.

In the paper, we present an elaborated survey of cyberbullying detection and prediction techniques and classify these different techniques using various factors used for analysis. The remainder of the paper is organized as follows: Sect. 2 presents the definition and background of cyberbullying. Section 3 is covering the related work done for cyberbullying detection techniques. Existing prediction techniques are discussed in Sect. 4. Section 5 presents the taxonomy of cyberbullying detection and prediction techniques. Section 6 is Comparative Analysis and discussion and we conclude in Sect. 7.

2 Background

Bullying is a very commonly observed act in adolescent. Bullying increased drastically in all age groups when it got the exposure to technology. In cyberspace, to bully someone, a person does not have to be physically strong. A person who has never bullied anyone in real life can also anonymously starts bullying on social media.

2.1 Statistics of Technology Utilization

The availability of electronic devices like laptops, i-Pads, smartphones, and their combination with Internet technology has become very easy. Internet made social networks very popular. All age group people are carrying these high-end devices and resultantly reachability of online social networks have increased. The recent survey

Table 1 Utilization of electronic devices by kids in the United Kingdom [5]

Age	Use of tablet (%)	Use of laptop or desktop (%)	Smartphones (%)
3–4	55	24	—
5–7	67	49	2
8–11	80	66	32

done by the Center for the Digital Future reports 92% of Americans are using the internet and, on an average, everyone is spending 23.6 h a week online [4]. The authors Kowalski et al. [5] have given statistics about kids in the United Kingdom using the electronic gadgets in the year 2016 (Table 1). Also, the authors state that in United States, 66% of elementary school youth are regularly using a laptop, whereas 53% are using smartphone, and 78% a tablet. Additionally, 35% of these youth owned a laptop, 36% owned a smartphone, and 69% owned a tablet. Application of Internet varies for different age groups but the common thing is the use of online social media for all. Lenhart, 2015 states about 71% of teens aged 13–17 have a profile on multiple social media platforms [6].

2.2 *Cyberbullying Definition, Types, Popularity in OSN and Outcome*

Definition: Traditional bullying is defined as an aggressive act that is intended to cause harm or distress, that is typically repeated over time, and that occurs among individuals whose relationship is characterized by a power imbalance [7, 8]. Based on this, many researchers have defined Cyberbullying. The author Giumentti et al. defined cyberbullying as rude/discourteous behaviors occurring through Information and Communication Technologies [9]. Cyberbullying is “an aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time against a victim who cannot easily defend him or herself” as defined by Smith and his colleagues [10].

Types of Cyberbullying: Cyberbullying can be of two types: direct or indirect cyberbullying [11]. Direct cyberbullying involves only two people the bully and a victim, whereas in indirect cyberbullying, a group of people can get involved. A very good example of indirect cyberbullying is a post on social media to make fun of someone and many people commenting and sharing it. Indirect cyberbullying has large and more dangerous impact. There are several categories of cyberbullying as stated by [12, 13]: Flaming, Masquerade, Denigration, Impersonation, Outing, Harassment, Trickery, Exclusion, and Cyberstalking. It is an act where people use bad words, sexting, online sexual solicitations, posting humiliating images or videos, tagging someone with the intention of making fun of him/her. Cyberbullying is possible through all different means but most common is OSN due to consistent availability of user and also due to the anonymity a person can achieve in cyberspace through

OSN. Not only teenagers but all age group people may suffer from the shameful act of cyberbullying.

Outcome of Cyberbullying: Cyberbullying can have very harmful effects on psychological health of a victim as well as on bully. The effect of cyberbullying is more harmful than traditional bullying because it can easily turn into indirect cyberbullying and victim cannot escape anywhere. Depression and isolation are majorly observed impacts of cyberbullying. D. Mann said emotional, concentration, and behavioral issues are reported by many victims [14]. These victims have also likely to reported frequent headaches, recurrent stomach pain, and sleeping difficulties. The survey done by the author Tjhin Wiguna et al. state that the extreme result of cyberbullying can be suicidal attempt [15]. Authors also state that impact of cyberbullying on male victims as well as perpetrators is, they become more aggressive and get addicted to alcohol consumption or cigarette smoking, whereas female victimization result in internalizing behaviors, such as ideation, isolation, depression, or suicide attempt. Research showed that higher level of cyberbullying results in higher level of depression [16]. The author Nixon also found that 32% of cyberbullying targets experienced at least one symptom of stress.

Through the social networks, many teenagers, as well as adults, are suffering from cyberbullying. Almost 50% of the youth in the United States have admitted being bullied [17]. But most of these victims usually hide their victimization due to different reasons. Teenagers or adolescents fear that the device may be taken away from them. Adults are embarrassed to acknowledge that they are being bullied, they also have a fear of being misunderstood by their peers or family members. Hence, the identification and reporting of cyberbullying is very important to find victims and take some action to cure them. It is also required to save society from the damage cyberbullying is causing.

3 Cyberbullying Detecting Techniques

Few social media sites are Facebook, Twitter, Instagram, LinkedIn, YouTube, Pinterest, Snapchat, ASKfm, etc. Out of these all social media platforms, Facebook, Twitter, YouTube, Ask.fm, and Instagram have been listed as top five networks with the highest percentage of users reporting experience of cyberbullying [18]. A huge amount of data is available and it goes on increasing every hour. This data is of different type like textual posts, comments, images, video, hashtags. Cyberbullying detection is possible by analyzing this content present on the social media. Most of the researchers are using machine learning approach for classifying the content in bullying and non-bullying categories. The classifiers can be binary or multiclass classifiers [19, 20]. Binary classifier classifies the content in 2 categories mostly positive and negative. Multi-class classifier can classify the data in multiple categories. Most of the researchers are focusing on binary classifiers.

3.1 Cyberbullying Detection Techniques Using Binary Classification

Cyberbullying identification is done based on the data present in different profiles of OSN. Most of the researchers are classifying the content in bullying and non-bullying categories. All of these methods are using some or other machine learning mechanism. To implement machine learning concept the major requirement is dataset for training and testing the machine. Major limitation is confined availability of datasets. Data scraping also provides restricted amount and type of data. According to a broad survey done by the authors Mahlangu et al. [21], the majority of researchers have generated their own datasets and some have scrawled websites. The authors [21] have also identified that for binary classification mostly SVM classifier is used by many researchers as it provides comparatively better results.

Different social media sites are focusing on different types of data, for example, Twitter is used for posting textual comments known as tweets, Instagram is used for sharing images, whereas Facebook contains combination of images, text, and video. Analyzing text is comparatively easy and focused by maximum researchers. Very few are talking about image content.

Profane Words-Based Detection: While considering textual data, one can simply focus on bad or insulting words identification. However, only the presence of bad word is not always bullying as stated by the authors Homa et al. [22]. Bullying is repeated aggressive act for hurting someone. Repeated behavior has to be checked before concluding bullying.

Author Sakshi Gujral have used a model considering only textual data and purely focusing on identification of bad words for detection as well as prediction of cyberbullying in presented paper [23]. Twitter media platform was selected and using twitter API and R studio, a balanced dataset was obtained containing bullying as well as non-bullying keywords. Tweets were treated as data after cleaning and preprocessing. Manual labeling is done as positive and negative based on sentiment. Bag-of-Words Approach with unigram feature extraction are used for extracting required features. Then by using term frequency feature reduction is done. CART, Decision Trees, Random Forest, Logistic Regression are used for prediction of cyberbullying. The authors have specified the proposed system but not mentioned internal details of implementation. The actual prediction is not clear. Naive Bayes classifier is used with modified version of Laplace function to detect the tweets potentially as bullied or not. Paper proposed a method specifically for twitter and only for text, where classification is done based on bad or negative word occurrence.

Yin et al. [24] used a supervised learning approach for detecting harassment. They used content, sentiment, and contextual features of documents to train the SVM classifier for a corpus of online posts. In this study, only the content of the posts was used to determine whether a post is harassing or not, and the characteristics of the author of the posts were not considered. Yin et al. [24] have used all the combination of these three features. In their study N-grams, TFIDF weighting and foul words frequency were used as baselines.

User Characteristics and Semantic Features-Based Detection: Most of the recent research papers focus on individual comments and do not take into account the context of the comment, user's comments history and user characteristics [25]. Authors Dadvar et al. hypothesized that gender-specific writing style and language features improve the overall detection accuracy [26]. There is a difference between the way male and female bully other people. Females generally use indirect tricks like excluding someone from group, whereas male openly use abusive or profane words, threatening expressions [27]. Argamon et al. found that females use more pronouns and males use more noun specifiers [28].

Maral Dadvar et al. have used SVM to train gender-specific text classifier. Data is discussion present on MySpace social media. Dataset consists of more than 381,000 posts in about 16,000 threads. Overall, 34% of the posts are written by female and 64% by male authors. dataset was manually labeled as harassing or non-harassing. Foul word analysis is done by comparing the most frequently used foul words used by each gender. As mentioned in [26], Maral Dadvar et al. have also used four types of features: profane words, second-person pronouns, other personal pronouns, and the weight of the words in each sentence, are frequently used for harassment classification. To evaluate the classification accuracy, they used tenfold cross-validation and calculated corresponding precision, recall and F-measure. Results of the paper proves gender-specific features improved the overall accuracy measures. Results are better in male-specific post than female-specific.

Cyberbullying comments mostly includes swear or insult word. Going beyond only text, the authors [29] are also identifying the association of swear and insult words with second person entity or person name to confirm bullying. The authors Yee Jang Foong et al. proposes validating association between these can encounter occurrence of cyberbullying. Author have explicitly evaluated the association between swear or insult word with second person entity. For achieving the goal, data from ASKfm is taken and processed. The approach is utilizing multiple features like tf–idf, LIWC and Dependency feature. The combination of each possible pair of these features have been employed. Labeling is done manually by using Amazon Mechanical Turk Service. Supervised learning is used and classification is done by SVM classifier. Only text messages are considered for analysis.

Though cyberbullying victims does not all belong to particular age range but as observed by authors Slonje and Smith; Williams and Guerra, cyberbullying behavior is highest among teenage users and it goes on decreasing as the age increases [30, 31]. Hence maximum researchers are focusing on cyberbullying detection for adolescents. Author Yasin and his colleagues, considering the same theory, have developed a parenting application named BullyBlocker for Facebook platform [32]. The model is specifically designed to be used by parents of teenagers. BullyBlocker intimate parents by sending an alert message when bullying occurs. The model is also built to measure the degree of cyberbullying by calculating bully rank. Bully rank estimates the probability of their child being bullied. As the application is developed for parents the major constraint is victim's login information is required for the application to run or to see the bullying results. It collects all recent wall posts, photos, comments,

and user profile information. Measuring vulnerability is calculated based on factors like moved to new school or new neighborhood. Age and gender features are also considered.

Authors [25] have not only focused on individual comments but also considered user characteristics and user comment history. A very rich feature set is used by the authors for dealing with false positive rate. Textual data, i.e., comments on YouTube videos are collected from YouTube media platform. Along with comments each comment's user id, date, and time was also stored. The final dataset of 4626 comments from 3858 distinct users was collected. Users with public profiles were stored to retrieve the comment history of last 6 months. On average 54 comments per user were stored in dataset. Labeling was done manually to all comments.

The rich feature set consists of 3 different features: Content-based features, Cyberbullying features and User-based features. In content-based feature traditional method of profane word identification is done and its association with first and second person pronoun is identified. Here authors also have considered emoticons normalized with text. In cyberbullying feature, topics, where bullying occurs, were selected, e.g., minority races, religions, and physical characteristics. Length of comments and use of capital letters are also considered as features to detect cyberbullying. Finally, in user-based feature, history of user comment is analyzed to identify the presence of offensive language used by user. This analysis helps to see the nature of user. 10fold cross-validation evaluated with precision, recall and F-measure was used to verify result. According to authors, profane words, Context-based information and user information have significant positive effect on the result. Capital letters do not add anything in significance. Age did have a positive impact but not very high. Length has no significant impact.

Social Features, Network Features, and Sentiment Features Based Detection: As the work in the field was getting enriched by new features and their positive impact on detection of cyberbullying, researchers did not remain focused on single feature. Considering multiple parameters to make detection more powerful was adapted by many researchers. The authors Michele et al. [33] have taken into account four important features: Syntactic, Semantic, Sentiment, Social. Textual sentences are analyzed for syntactic and semantic features. The system is designed for Twitter data and later also tested for YouTube and Formspring. The novelty of the method is Unsupervised approach which other researchers have not considered, which avoids manual labeling and hence makes the system more automatic. Assumption of authors is cyberbullying post is extremely negative. Under Syntactic features, the authors have targeted bad words identification, their density, Density of uppercase letter in a sentence and presence of Exclamations and Question marks. Semantic feature only focused on the presence of personal pronoun with bully word. Bigram and trigram are used for this. Semantic feature can be more explored. Sentiment polarity is identified in terms of positive and negative. Emoticons are also used for calculating sentiment polarity. Social feature checks if there is any direct user tagging with profane word. It proposes to measure politeness of the user sending post but for this complete

comment history of the user is required. Unsupervised learning is achieved using Self-Organizing Maps.

Authors Mohammed Ali et al. [34] have developed a model for detecting cyberbullying in Twitter. Features considered are user's features from tweets, such as network, activity, user, and tweet content. The model used machine learning classifiers to classify the tweets into two categories cyberbullying or non-cyberbullying. For feature selection, three algorithms were selected by authors: c2 test, information gain, and Pearson correlation. The selected classifiers are NB, LibSVM, random forest, and KNN. Geo-tagged tweets were collected from tweeter containing: user ID, username, user biography, user screen name, user URL, user account creation data, tweet text, tweet creation time, tweet's unique ID, language of tweet, number of tweets of a user, number of favorites, number of followers, number of mentions, number of following, number of retweets, bounding box of the location (geolocation), and the application that sent the tweet.

Multilingual Cyberbullying Detection: Cyberbullying detection modules are designed for many languages other than English, e.g., Japanese, Turkish, Arabic. Natural Language Processing is used for processing the collected multilingual data. The authors Haidar et al. [19] claims that they have considered Arabic language for the first time to detect cyberbullying and in [17] they have proposed a system for detecting cyberbullying in English and Arabic language using machine learning and NLP concepts. WEKA toolkit is used, as it supports Arabic language. Data was scrapped from Twitter and Facebook and then cleaning and preprocessing was done. Textual feature is only considered in this paper. SVM and Naive Bayes classifiers were used. Through the paper authors were able to prove detection of cyberbullying is possible in Arabic language. The author also have discussed about all binary classification methods like Naive Bayes, nearest neighbor, SVM, and decision tree for classifying the text in bullying or non-bullying.

Another system developed on multilingual basis is by authors Michal P. et al. They worked on the detection of cyberbullying in Japanese language. The authors state [35] Japan Parent–Teacher association started and activity called “Internet-patrol” for detecting harmful entry whenever entered and ask administrator to remove it. This monitoring was done manually hence was taking too much time and manual efforts. The authors [35] have proposed a method to automate the classification of all sentences from the dataset into harmful and non-harmful entries. A seed word is considered and then every harmful word is categorized into three categories: obscene, violent and abusive. Maximal relevance value for each seed word with words contained in the input entries from each category is calculated. Sometimes even if a word in itself is not harmful, it gains harmful meaning when used in a specific context, or in combination with other words. Considering this, the method is extended to calculate relevance score for each phrase automatically. Data is collected from electronic bulletin board pages. The authors have contributed with a very important observation that if the same model is tested after a year, over 30% performance drop is observed and hence it should keep on updating.

3.2 *Detection Techniques Using Multiclass Classification*

Very few researchers have used multiclass classification. Multiple classes can be based on the severity of bullying or topic of discussion like race, body structure or sexual solicitations. By analyzing sentiment research can contribute to identifying different emotions behind harassing comments and their responses to categories bullying content in multiple categories. These different categories can also be identified by studying the posted images.

Profane Words Based Detection: The framework [36] uses multiple classes. Existence of cyberbullying is measured from non-cyberbully to severe level of cyberbullying by gradually incrementing the level of severity. It uses various classifiers like SVM, Naive Bayes, Linear, Poly, RBF, and sigmoid kernels based and comparative analysis is provided. The most optimal SVM kernel in classifying cyberbullying is the Poly kernel with an average accuracy of 97.11%. Textual conversation in Formspring.me is taken from Kaggle to form datasets. Question and answer type of data is present in Formspring.me. 1,600 conversations are divided into training and testing sets. Collected data is imported to Rapid Miner for further processes: Preprocessing, Extraction, Classification, and Evaluation. Labeling is done manually before the process of classifying text into different severity based bullying classes.

Often machine learning is not sufficient to classify text into multiple categories. Authors [37] have used Fuzzy logic and Genetic algorithm for classifying the presence of cyberbullying into different types of activities: Flaming, Harassment, Racism, and Terrorism. Based on the category of bullying action can be taken. Fuzzy rules are used for classification and genetic algorithm is used for optimizing the parameters and to obtain precise output. Data is collected from formspring.me and preprocessing is done to remove hyperlinks, extra characters and stop words. Features like Noun, Adjective, and Pronoun from the text and frequency of occurrence in the text are extracted and given to learning algorithm unit which is composed of genetic algorithm. Knowledge extracted is then passed to fuzzy rule set. This knowledge is kept in a population of chromosomes, which is processed by the genetic algorithm. The output from learning unit is given to classifier technique which classifies the cyberbully activities using the fitness value of chromosome. Paper gives a systematic algorithm and experimental results showing accuracy of the system.

User Characteristics Based Detection: User personality, is one of the major user characteristics. Empirical studies have shown links between an individual's personality with both cyberbullying and traditional bullying [38–40]. Personality is a characteristic set of behaviors that distinguishes one person from another. Studies have proved that personality of a user can be automatically predicted based on his/her online communication style. In a very recent study, the authors [41] have presented a model for cyberbullying detection using user personality identification. User personality is determined by using Big Five and Dark Triad models. The user is categorized into four types: bully, aggressor, spammer, and normal. Data is collected from twitter. The results in the paper shows that the impact of considering user's personality is, the improvement in performance of cyberbullying detection model, which is achieving

up to 96% precision and 95% recall. The authors have used a completely new concept and investigated if user's personality can be used for automatic cyberbullying detection on their textual communication using artificial intelligence. Random forest algorithm has been used for above.

4 Cyberbullying Prediction Techniques

Prediction of attacks is always better than detection after it occurs. Researchers have started research in the direction of predicting cyberbullying using the existing data present on OSN. Few prediction techniques are present which are dealing with text content only. Images and videos are very less explored [21].

One of the recent papers [22] deals with Instagram social media where classification of images in bullying and non-bullying is done by authors Homa HosseiniMardi et al. The dataset consists of 3165 K media sessions. Media sessions include images and their associated comments. 25 K user profiles details were also stored which included the images and their comments users have posted, user id of each follower and following and user id of each who have commented or liked a post of user. Labeling was done manually by CrowdFlower website, where only media session with some profanity was passed for labeling to reduce the labeling cost.

Authors [22] claim that all the previous work is actually only detecting cyber-aggression or mere negativity. Cyberbullying is a repeated act or aggression and hence the frequency of aggression plays a very vital role in cyberbullying detection. To detect the repetition, media session with minimum 15 comments were selected. Features authors are considering are profanity, social graph, temporal commenting behavior, linguistic content, and image. Experimental results in the paper proves that only the use of profane words does not mean cyberbullying. Cyberbullying is a subset of cyberaggression. Going beyond cyberaggression is really required.

For prediction data, labeling was done for the mixed type of data where comments with and without profanity was considered. Votes are calculated to identify the confidence level and low confidence level is considered as prediction of cyberbullying. A completely new direction is given by the author [22] but more or less things are not automated. Prediction mechanism is introduced but it is based on manual labeling only. For prediction, many more features can be considered. Image recognition part is manual and sarcasm is not considered.

Another paper claiming automated prediction of cyberbullying by the same authors have also considered Instagram social media [42]. The prediction is dependent on initial posting of the media object, i.e., image with caption followed by comments. Authors have considered text caption, image content, as well as social graph parameters and temporal content behavior. In social graph, parameter followers and followings are mapped and temporal parameter includes post time. To design and train the classifier, fivefold cross-validation was applied to the data. Logistic regression classifier has been applied to train a predictor with the forward feature selection approach.

5 Taxonomy of Cyberbullying Detection and Prediction Techniques

Literature review consists of multiple methods. Different researchers are working and trying to explore all different angles of cyberbullying detection. We have arrived at the following taxonomy, after considering all the techniques studied in literature survey. This taxonomy gives a clear picture about the various proposed methods of cyberbullying detection. It also shows in which area work is going on and also the area which is unexplored or less explored. Taxonomy is as follows (Fig. 1).

From the study, we could find the user character is an important aspect in detecting and predicting any cybercrime. Even for Cyberbullying detection, the user account plays an important role. It becomes easy for perpetrators to hide behind fake identity and bully anyone. The Parameter “User profile legitimacy” shown in dotted red box can be the key point for cyberbullying prediction. This part is not yet considered by the researchers for prediction of cyberbullying and we will be targeting it for our further research.

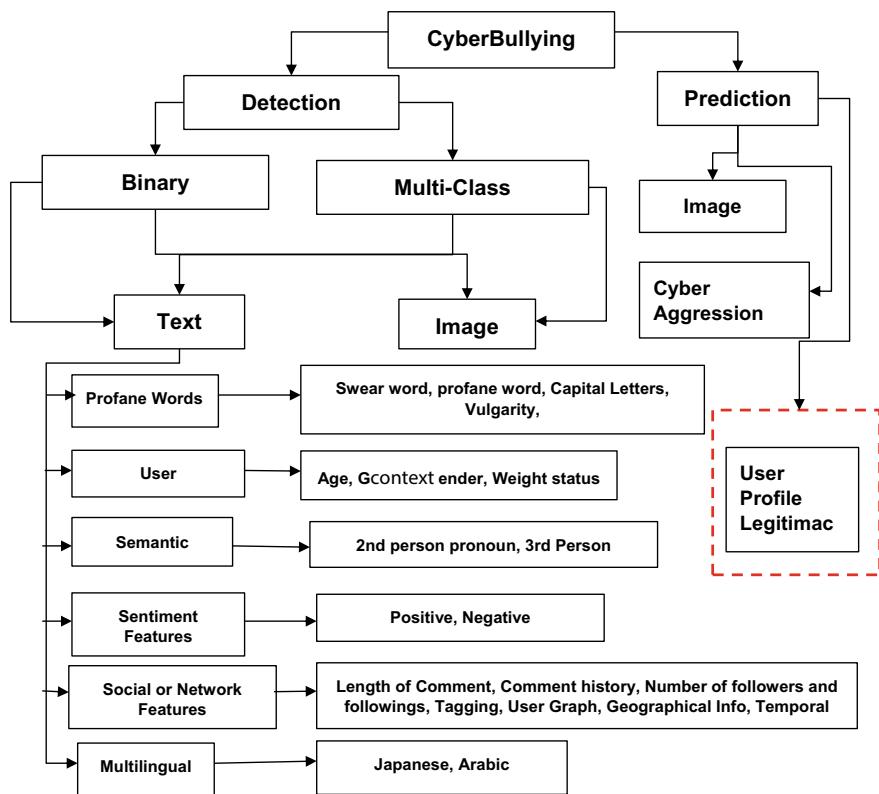


Fig. 1 Taxonomy of cyberbullying detection and prediction techniques

6 Analysis

After studying all different methods, we have done a comparative analysis based on some important features such as The type of classification, features considered to detect cyberbullying, size of dataset and social networking site, machine learning algorithm, and presence of prediction. Many Methods are using SVM classifier as the research shows it provided good results. Only one paper talks about unsupervised approach as it is relatively more complicated (Table 2).

From the classification and taxonomy, it can be observed that very less work is focusing on image analysis for detection of cyberbullying occurrences. The major problem is limited availability of datasets. Many researchers are creating and labeling datasets on their own. Every researcher is using one or couple of features leaving remaining features unattended, which affect the accuracy of detection. Considering maximum features may improve the detection as well as it will also help in finding the severity of cyberbullying.

Use of sentiment analysis is limited to only positive and negative sentiments by the researchers. Sentiment analysis may also lead to decrease the false positive rate in the detection of cyberbullying by identifying the depth of emotion used in textual content. It can also be used for differentiating cyberaggression from cyberbullying.

Cyberaggression is well differentiated from cyberbullying by authors Homa Hosseiniardi et al. [22]. Cyberaggression can be used as indication that cyberbullying may occur in the future. User characteristics also includes the legitimacy of the user profile. Most of the crimes are done using cloned or fake profile. Bully can also hide himself behind false names [33]. No authors have yet considered this feature for detection or prediction of cyberbullying as per our survey. This is, therefore, our future scope of work. Authors Patxi et al. [20] has proposed a system to detect troll profiles and later it is applied to a real-life cyberbullying in elementary school. Identifying the legitimacy of profile will help in automated cyberbullying prediction.

7 Conclusion

After comparing the recent work in cyberbullying detection and prediction we observed that most of the researchers are focusing on those social media for which datasets are easily available. Popular social media platforms are still not secure in aspect of cyberbullying. Detection accuracy can be improved by considering multiple parameters with which cyberbullying is associated. Very few researchers have actually achieved prediction of cyberbullying. We understand that the first step toward prediction is identification of cyberaggression. User profile legitimacy detection will definitely provide a great help in prediction of cyberbullying in online social networks.

Table 2 Comparative analysis of cyberbullying detection and prediction techniques

	Classifier binary/multiclass	Feature	Social media platform	Size of dataset	Machine learning technique	Prediction
Hectoring detector [23]	Binary	Text	Twitter	Not specified	Bag of Words	Yes
Supervised harassment detector [24]	Binary	Text, semantic/context	Kongregate, Slashdot and MySpace	Around 10 K posts	TFIDF	No
Gender-specific detector [26]	Binary	Text, gender information	MySpace	381,000 posts	SVM	No
Profane word with 2nd person association [29]	Binary	Text, semantics	ASKfm	10,000 questions and answers	SVM	No
BullyBlocker [32]	Binary	Text, user char., social features	Facebook	All posts of one profile	Not specified	No
User context based detector [25]	Binary	Text, temporal, user char., network, syntactic	YouTube	4626 comments, 3858 user profiles	10fold cross-validation	No
Unsupervised detector [33]	Binary	Syntactic, semantic, sentiment, social	Twitter, YouTube, and Formspring	Not specified	Self-organizing maps	No

(continued)

Table 2 (continued)

	Classifier binary/multiclass	Feature	Social media platform	Size of dataset	Machine learning technique	Prediction
Online cybercrime detector [34]	Binary	Network, activity, user, and tweet content	Twitter	2.5 million geo-tagged tweets	NB, SVM, decision trees (random forest), and KNN	No
Multilingual cyberbullying detector [17]	Binary	Profane words	Facebook, Twitter	35273 Arabic tweets, 9431 English tweets	SVM and Naive Bayes	No
Cyberbullying severity detector [36]	Multiclass	Text (profane words)	Formspring.me	1,600 conversations	SVM, Naive Bayes, Linear, Poly, RBF, and sigmoid kernels based	No
Intelligent cyberbullying detector [37]	Multiclass	Syntactic, semantic	Formspring.me, Myspace	Not specified	Fuzzy logic and genetic algorithm	No
User personality based detector [41]	Multiclass	User personality	Twitter	9484 tweets	Random forest	No
Cyberbullying Predictor [22]	Binary	Text (profane words), image, social graph	Instagram	3165 K media sessions, 25 K user profiles details	SVM, Logistic regression	Yes
Automated prediction of cyberbullying [42]	Binary	Image, profane words, social graph, temporal content	Instagram	1,164 media sessions, 25 K public user profiles	Logistic regression	Yes

References

1. D. Boyd, N. Ellison, Social network sites: definition, history, and scholarship. *J. Comput. Med. Commun.* **13**, 210–230 (2007)
2. Z. Shan, H. Cao, J. Lv, C. Yan, A. Liu, Enhancing and identifying cloning attacks in online social networks, in *Proceedings of the 7th International Conference on Ubiquitous Information Management and Communication*, Article No. 59 (ACM, Kota Kinabalu, Malaysia, 2013), pp. 17–19
3. M. Di Capua, et al., Unsupervised cyber bullying detection in social networks, in *23rd International Conference on Pattern Recognition (ICPR)*, México, 4–8 Dec 2016)
4. Digital Future Project Survey 2017, <http://www.digitalcenter.org/wpcontent/uploads/2013/10/2017-Digital-Future-Report.pdf>
5. R. Kowalski, S.P. Limber, A. McCord, A developmental approach to cyberbullying: prevalence and protective factors. *Artic. Aggress. Violent Behav.* **45**, 20–32 (2019) (Elsevier)
6. A. Lenhart, Teens, social media & technology overview 2015 (Internet American Life Project). Pew Research Center, Aug 2015, <http://www.pewinternet.org/2015/04/09/teens-social-media-technology-2015>
7. D. Olweus, *Bullying at School: What We Know and What We Can Do* (Blackwell, New York, 1993)
8. D. Olweus, School bullying: development and some important challenges. *Annu. Rev. Clin. Psychol.* **9**, 1–14 (2013). <https://doi.org/10.1146/annurev-clinpsy-050212-185516>
9. G.W. Giumenti, E.S. McKibben, A.L. Hatfield, A.N. Schroeder, R.M. Kowalski, Cyber incivility @ work: the new age of interpersonal deviance. *Cyberpsychology Behav. Soc. Netw.* **15**, 148–154 (2012). <https://doi.org/10.1089/cyber.2011.0336>
10. P.K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, N. Tippett, Cyberbullying: its nature and impact in secondary school pupils. *J. Child Psychol. Psychiatry* **49**(4), 376–385 (2008). <https://doi.org/10.1111/j.1469-7610.2007.01846.x>
11. C. Langos, Cyberbullying: the challenge to define. *Cyberpsychology Behav. Soc. Netw.* **15**(6), 285–289 (2012). <https://doi.org/10.1089/cyber.2011.0588>
12. N. Willard, *Educator's Guide to Cyberbullying and Cyberthreats*. Center for Safe and Responsible Internet Use (2007)
13. N. Samanah, A. Masrah, M. Azmi, M.S. Nurfadhilna, A. Mustapha, S. Shojaee, A review of cyberbullying detection. An overview, in *13th International Conference on Intelligent Systems Design and Applications (ISDA)* (2013)
14. D. Mann, Emotional Troubles for ‘Cyberbullies’ and Victims. WebMD Health News, 6 July 2010. <http://www.webmd.com/parenting/news/20100706/emotional-troublesfor-cyberbullies-and-victims>. Accessed 24 Aug 2015
15. T. Wiguna, I.R. Ismail, R. Sekartini, N.S.W. Rahardjo, F. Kaligis, A.L. Prabowo, R. Hendarmo, The gender discrepancy in high-risk behaviour outcomes in adolescents who have experienced cyberbullying in Indonesia. *Asian J. Psychiatry* **37** (2018) (Elsevier)
16. C. Nixon, Current perspectives: the impact of cyberbullying on adolescent health, in *Adolescent Health, Medicine and Therapeutics* (2014), p. 143
17. B. Haidar, M. Chamoun, A. Serhrouchni, Multilingual cyberbullying detection system, detecting cyberbullying in Arabic content, in *1st Cyber Security in Networking Conference (CSNet)* (IEEE, 2017)
18. Ditch the Label Anti Bullying Charity, The annual cyberbullying survey 2013 (2013), <http://www.ditchthelabel.org/annual-cyber-bullying-survey-cyber-bullying-statistics/>
19. B. Haidar, M. Chamoun, F. Yamout, Cyberbullying detection: a survey on multilingual techniques, in *European Modelling Symposium (EMS)* (IEEE, 2016)
20. P. Galán-García, J.G. Puerta, C.L. Gómez, I. Santos, P.G. Bringas, Supervised machine learning for the detection of troll profiles in twitter social network: application to a real case of cyberbullying. *Log. J. IGPL* **24**(1) (2016)

21. T. Mahlangu, C. Tu, P. Owolawi, A review of automated detection methods for cyberbullying, in *International Conference on Intelligent and Innovative Computing Applications (ICONIC)* (IEEE, 2018)
22. H. HosseiniMardi, R.I. Rafiq, R. Han, Q. Lv, S. Mishra, Prediction of cyberbullying incidents on the Instagram social network, in *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (IEEE/ACM, 2016)
23. S. Gujral, Predicting and detecting hectoring on social media using machine learning, *Int. J. Comput. Sci. Eng.* **5**(8) (2017)
24. D. Yin, Z. Xue, L. Hong, B.D. Davison, A. Kontostathis, L. Edwards, Detection of harassment on Web 2.0, in *Proceedings of the Content Analysis in the WEB 2.0 (CAW2.0) Workshop at WWW2009*, Madrid, Spain (2009)
25. M. Dadvar1, D. Trieschnigg, R. Ordelman, F. de Jong, Improving cyberbullying detection with user context, in *European Conference on Information Retrieval ECIR: Advances in Information Retrieval* (Springer, 2013), pp. 693–696
26. M. Dadvar, F.D. Jong, R. Ordelman, D. Trieschnigg, Improved cyberbullying detection using gender information, in *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop* (2012)
27. J.F. Chisholm, Cyberspace violence against girls and adolescent females. *Ann. N. Y. Acad. Sci.* **1087**, 74–89 (2006)
28. S. Argamon, M. Koppel, J. Fine, A.R. Shimoni, Gender, genre, and writing style in formal written texts. *Text Interdiscip. J. Study Discourse* **23**, 321–346 (2003)
29. Y.J. Foong, M. Oussalah, Cyberbullying system detection and analysis, in *European Intelligence and Security Informatics Conference* (IEEE, 2017)
30. R. Slonje, P.K. Smith, Cyberbullying: another main type of bullying? *Scand. J. Psychol.* **49**(2), 147–154 (2008)
31. K.R. Williams, N.G. Guerra Prevalence and predictors of internet bullying. *J. Adolesc. Health* **41**(6), S14–S21 (2007)
32. Y.N. Silva, C. Rich, D. Hall, BullyBlocker: towards the identification of cyberbullying in social networking sites, in *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (IEEE/ACM, 2016)
33. M. Di Capua, E. Di Nardo, A. Petrosino, Unsupervised cyber bullying detection in social networks, in *23rd International Conference on Pattern Recognition (ICPR)*, Cancún Center, Cancún, México, 4–8 Dec 2016)
34. M.A. Al-garadi, K.D. Varathan, S.D. Ravana, Cybercrime detection in online communications: the experimental case of cyberbullying detection in the Twitter network. *J. Comput. Hum. Behav.* **63**, 433–443 (2016) (Elsevier)
35. M. Ptaszynski, F. Masui, T. Nitta, S. Hatakeyama, Y. Kimura, R. Rzepka, K. Araki, Sustainable cyberbullying detection with category-maximized relevance of harmful phrases and double-filtered automatic optimization. *Int. J. Child Comput. Interact.* (2016) (Elsevier)
36. Novianto, S.M. Isa, L. Ashanti, Cyberbullying classification using text mining, in *1st International Conference on Informatics and Computational Sciences* (IEEE, 2017)
37. B. Sri Nandhinia, J.I. Sheebab, Online social network bullying detection using intelligence techniques, in *International Conference on Advanced Computing Technologies and Applications* (2015) (Elsevier)
38. S. Resett, M. Gamez-Guadix, Traditional bullying and cyberbullying: differences in emotional problems, and personality. Are cyberbullies more Machiavellians? *J. Adolesc.* **61**, 113–116 (2017)
39. M. van Geel, A. Goemans, A. Toprak, P. Vedder, Which personality traits are related to traditional bullying and cyberbullying? A study with the Big Five, Dark Triad and sadism. *Pers. Individ. Differ.* **106**, 231–235 (2017)
40. R. Festl, T. Quandt, Social relations and cyberbullying: the influence of individual and structural attributes on victimization and perpetration via the Internet. *Hum. Commun. Res.* **39**(1), 101–126 (2013)

41. V. Balakrishnana, S. Khana, T. Fernandezb, H.R. Arabniac, Cyberbullying detection on twitter using Big Five and Dark Triad features. *J. Pers. Individ. Differ.* **141** (2019) (Elsevier)
42. H. HosseiniMardi, S.A. Mattson, R.I. Rafiq, R. Han, Q. Lv, S. Mishr, Prediction of cyberbullying incidents in a media-based social network, in *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (IEEE/ACM, 2016)

A Formal Modeling Approach for QOS in MQTT Protocol



E. Archana, Akshay Rajeev, Aby Kuruvila, Revathi Narayankutty and Jinesh M. Kannimoola

Abstract With the rising demand for IoT devices, communication protocols like MQTT, CoAP, and many more, have become an integral part of the system to ensure safe and reliable data transfer. Using lightweight communication protocols such as the Message Queuing Telemetry Transport (MQTT) protocol makes it much easier to establish communication between distributed devices as it easily recovers from connectivity loss, component failures, and loss of packets. The pivotal contribution of this paper is the method of approach to formally model, analyze, and verify the Quality of Service (QoS) levels of the MQTT protocol. A complete analysis of the Quality of Service levels is performed to confirm that it behaves correctly as specified when used in communication between different components. Formal modeling has been done using PROMELA language and the model verification is done using a system verification tool called SPIN Model Checker.

Keywords IoT · MQTT · Formal verification · PROMELA · SPIN · Model checking

E. Archana · A. Rajeev · A. Kuruvila · R. Narayankutty (✉)

Department of Computer Science and Engineering, Amrita School of Engineering,
Amrita Vishwa Vidyapeetham, Amritapuri, India

e-mail: revathynarayankutty@gmail.com

E. Archana

e-mail: archanae21@gmail.com

A. Rajeev

e-mail: akshayrajeev@gmail.com

A. Kuruvila

e-mail: abykuruvila3941@gmail.com

J. M. Kannimoola

Department of Computer Science and Applications, Amrita School of Engineering,
Amrita Vishwa Vidyapeetham, Amritapuri, India

e-mail: jinesh@am.amrita.edu

1 Introduction

The number of devices used by an individual multiplies as the world rapidly develops. Internet of Things (IoT) has become widely popular as it helps these devices to easily connect and manage from anywhere, at any time. It provides a medium for devices to connect, interact, and exchange data [1]. Thus, it has become a well-established fact that IoT brings technology close to people [2, 3]. Due to its simplicity, IoT has become prevalent in various industries like health care, automotive, manufacturing, power grid, domestics, and many more [4]. IoT has been identified as a potential solution to alleviate the pressure on health care systems and it becomes a leading topic of research in the IoT field. Examples include remote monitoring of patients with conditions like Parkinsons and diabetes. Failure of such a system can be devastating and hence, all measures to prevent such a scenario should be taken.

Since there is a huge amount of critical data transfer between devices, it is absolutely necessary to have reliable communication protocols to handle the traffic in mission-critical application [4, 5]. Transferring such sensitive data can open up new security challenges that can cause catastrophic results [6]. Thus, it is important to understand and verify various IoT communication protocols, such as eXtensible Messaging and Presence Protocol (XMPP), Advanced Message Queuing Protocol (AMQP), Message Queuing Telemetry Transport Protocol (MQTT), and The Constrained Application Protocol (CoAP) [7].

This research paper focuses on the Message Queuing Telemetry Transport Protocol (MQTT), which is a standard publish and subscribe messaging transport protocol that works on top of the TCP/IP protocol [8]. MQTT helps in loss-less distribution of a stream of bytes in both directions in an ordered manner. It has numerous features that include [9]:

- Use of the publish/subscribe message pattern for message distribution among different clients.
- A messaging transport that is a skeptic to the content of the payload.
- Three quality of service levels for message delivery.
- Notifies the clients on the occurrence of ungraceful disconnections.

In this paper, we propose a formal model and verification of the Quality of Service (QoS) levels of the MQTT protocol [4]. Quality of Service levels defines message delivery guarantee between senders and receivers of a message, effectively it defines the reliability of communication. It is important to verify this feature to ensure that the messages are transferred correctly among the clients. Formal verification uses mathematical techniques to prove the correctness of a system. It also provides a systematic way to discover protocol flaws. Formal verification of software programs is done to prove that a program satisfies a intended behavior [3]. It is to check if the product or the program conforms to the specifications. In our approach, the protocol feature is modeled in PROMELA languages and verified using a model checker tool called SPIN [10]. The Linear time Temporal Logic (LTL) statements are used

to specify the correctness properties. Model checking is a method to exhaustively verify the correctness of a system according to its specifications.

The paper is organized as follows. Section 2 discusses the existing work in formal verification of various IoT communication protocols. Section 3 introduces the Quality of Service feature, SPIN Model Checker, PROMELA language, and mainly highlights the formal modeling and verification of the QoS levels. Section 4 presents the conclusion and future works.

2 Related Works

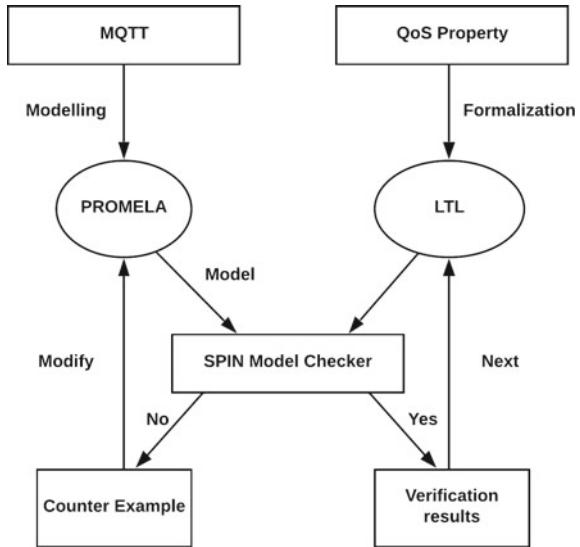
This section will be discussing the existing works on formal verification of IoT communication protocols. Vattakunnel et al. [10], proposes a verification model for application layer protocols. Intercommunication of layers were integrated into the research to correctly model the overall behavior of the system. The concept illustrated in the paper is formal verification of the Constrained Application Protocol (CoAP). Message exchanges among diverse clients are modeled in this research, along with the verification of its safety properties. The verification results were scrutinized based on memory usage, state transitions, and the search depth attained. The paper elaborates on the validation model for a multi-hop topology, which was built using PROMELA language. Further, the properties were verified using SPIN model checker. The methodology proposed in this work is efficient to verify any application layer protocol in IoT that run on top of the routing layer.

Chouali et al. [4] put forward methods to conquer the drawbacks of the MQTT protocol in communicating vehicles. The issue was the reliability of the protocol to transfer massive volume of data from the vehicles through a broker. Thus, a variant of MQTT protocol named MQTT-CV was proposed which processes the data sent by the broker. MQTT-CV was formally analyzed to ensure the authenticity of the protocol with respect to the security properties. The paper proved the correctness of MQTT- CV as it satisfies the property related to deadlock states and also proved that the vehicles behave correctly without any deadlocks.

Concerning the security properties of the MQTT protocol, Aziz et al. [8], projects a formal approach to prove that in various scenarios QoS requirements are not fulfilled. A framework using Event-B was proposed by Diwan et al. [1], which were used to model IoT protocols like MQTT- SN, CoAP, and MQTT. They have formally verified properties related to the persistent session, retained messages, will, and resource discovery. Based on the latest specifications of the MQTT protocol, Tena et al. [7] have presented a formal CPN model. Simulations were conducted for the validation of the model.

In our approaches, we are verifying the Quality of Service (QoS) levels of MQTT protocol. QoS is important and is a necessary factor that guarantee reliability, security, and quality of the communication. The respective feature is modeled in PROMELA and its verification is done using a powerful model checking tool, SPIN.

Fig. 1 Verification of system design



3 Formal Modeling

Formal verification is important to ensure the correctness of a system. PROMELA language is used to build the verification model of the QoS levels, which is then verified using the SPIN Model Checker.

As per Fig. 1, the system to be built, that is the Quality of Service levels of the MQTT protocol is modeled in PROMELA. In PROMELA, concurrent processes can be created dynamically and communication via message channels can be defined to be synchronous (not stored or buffered) or asynchronous (buffered). The properties to be verified are specified using Linear Temporal Logic (LTL). The model is then fed to SPIN model checker for verification. If the property that has been defined does not match with the logic of the validation model designed, then the model checker provides a counterexample. A counterexample is the path of execution that is tracked when a property fails. Therefore according to the counterexample, the model is modified and verified again until no more errors occur.

3.1 SPIN Model Checker and PROMELA

SPIN (Simple Promela INterpreter), an open-source model checker, developed at the Bell Labs by Gerard Holzmann, is one of the most powerful and popular tools that detects software defects in system designs. It is written in ANSI standard C and is portable across multiple platforms. Concurrent systems are specified in the modeling language called PROMELA.

PROMELA (Process or Protocol Meta Language) is a process specification language that is close to C language (in data types and declaring variables). It is a nondeterministic and guarded command language. In this model:

- The **processes** are global objects, which specify the behavior, communicate over channels and shared variables and is executed asynchronously. The keyword used to define the process is *proctype*.
- The **message channels** are synchronous or asynchronous channels. It does inter-process communication and the keyword used is *chan*.
 - **Sending the message (!)**
ch! 0 - sending over channel ch
 - **Receiving message (?)**
ch? c - receives from channel ch and pass to c.
- The **variables** are local and global and the basic data types used are *int*, *byte*, *mtype*, *bit*, *bool*.

More details about the grammar in PROMELA is available in [11].

3.2 Quality of Service (QoS)

Application messages in MQTT are delivered according to the Quality of Service (QoS) levels. A message can be delivered using one of the three QoS levels or a combination of it. The delivery protocol is very symmetric, and the clients can behave as senders, receivers, or brokers according to its requirements. Every client with whom the broker communicates are treated independently, as the protocol solely focuses on one-to-one communication.

For clients to communicate with each other, they must be connected to the same topic. The topics are created by a subscribing or publishing client, and they are not permanent. Topics are created to avoid uninterested clients from getting the messages passed through the channel. Once the topic has been created and the clients have subscribed to it, message passing can begin via the broker. The client always publishes a message to the broker and the broker transfers the message to all the respective subscribers.

There are three Quality of Service levels for message delivery, namely:

- At most once (QoS = 0): Messages are sent only once and message loss can occur.
- At least once (QoS = 1): Duplicate messages can occur but message arrival is assured at least once.
- Exactly once (QoS = 2): Message arrival is exactly once and it is assured.

The client always subscribes to a topic by specifying a certain QoS level. Later as the broker transfer the message to a subscriber, it uses the QoS level of the subscription made by that client. Hence, multiple QoS levels can be used in a single message delivery. For example, the sender publishes a message using QoS = 2, but if the

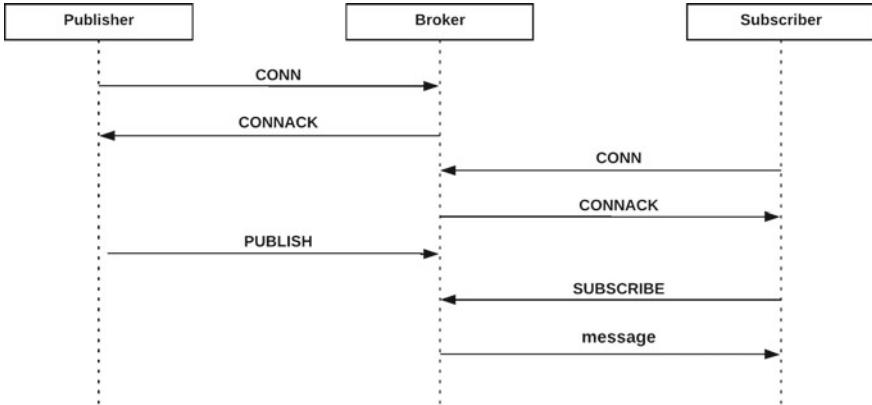


Fig. 2 Sequence diagram of QoS 0

receiver has subscribed to a topic using $\text{QoS} = 0$, then the receiver will proceed with its $\text{QoS} = 0$ itself.

3.3 Formal Modeling of Quality of Service Levels

There are three QoS levels for the MQTT protocol. Modeling and verification of these three levels are illustrated below.

3.3.1 QoS 0 (At Most Once)

$\text{QoS} = 0$, commonly known as “fire and forget” is the lowest QoS level. This service assures best-effort delivery of messages and no guarantee when or if the message will be delivered. The receiver of the message does not send any kind of acknowledgment if the message is received from the sender.

We have modeled the level considering its basic property to enable communication between clients. As shown in Fig. 2, to establish communication, we have created three clients: publisher, subscriber, and broker. The subscriber is the receiver who receives all the messages when it has subscribed to a particular topic that was created by the publisher. The broker is a medium through which both the clients communicate. Publishers and subscribers can interchange their roles depending on their needs.

```

init {
  run publisher(2, 2, 2, 0, 5);
  run subscriber(3, 3, 2, 0, 2);
  run publisher(0, 0, 1, 0, 1);
}
  
```

```

run subscriber(1, 1, 1, 0, 0);

}

```

The verification model that we have created works as follows:

- The clients send a connection request along with its client ID to the broker and wait for an acknowledgment.
- The broker upon receiving the connection request and client ID sends the acknowledgment message and the same ID back to the respective client.
- Once the connection is established, the client who acts as the publisher sends a *PUBLISH* packet to the broker. The *PUBLISH* packet will contain the client ID, packet ID, topic name, message, and the QoS level that it uses.

```

publisher!PUBLISH, client_id,
packet_id, topic_name, message,
qos_type;

```

- Another client who acts as the receiver sends a *SUBSCRIBE* packet to the broker. The *SUBSCRIBE* packet will have a client ID, packet ID, topic name, and the QoS level that it uses. The publisher and subscriber can use different QoS levels.

```

subscriber!SUBSCRIBE, client_id,
publisher_id, packet_id,
topic_name, qos_type

```

- Now, the *PUBLISH* packet and the *SUBSCRIBE* packet is with the broker. The broker then extracts the topic name from both the packets to compare. If the subscriber has subscribed to the same topic name as the one which was created by the publisher, then the broker sends the message of the publisher to the subscriber.

```

topic_name_sub == topic_name_pub
if
::qos_type_sub == qos_type_pub ->
send_message!message
fi

```

The clients communicate with each other though promela channels.

```

chan publisher = [0] of {mtype, short, short,
short, short, short};
chan subscriber = [0] of {mtype, short, short,
short, short, short};

```

A *timeout* statement is given so that the clients do not have to wait indefinitely to get the connection acknowledgment from the broker. Once it exceeds the time limit, the client will send a connection request again.

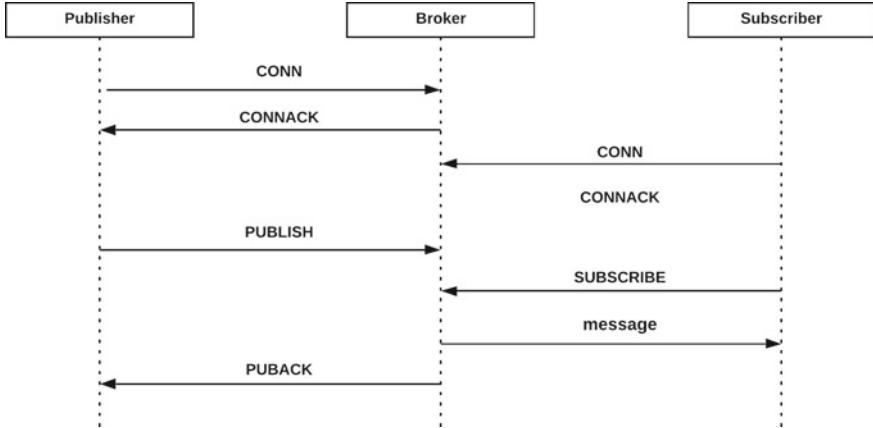


Fig. 3 Sequence diagram of QoS 1

```
(timeout == true) }-> {goto connect}
```

The system is modeled in such a way that multiple clients can communicate with the broker. Since QoS 0 is at most once, the subscriber may get the message once or never. Thus, there is no guarantee for message delivery. The model is checked using SPIN model checker. As a result, we get state-space diagram for the processes we have created in the model.

3.3.2 QoS 1 (At Least Once)

In QoS 0, there is no guaranteed message delivery among clients. This issue can be solved by using QoS 1, as every delivered message packet gets an acknowledgment back (refer Fig. 3). It guarantees that a message is delivered at least once to the subscriber but there is also a possibility for same messages to be sent or received multiple times.

```

init{
    run subscriber(10)
    run subscriber(10)
    run subscriber(20)
    run publisher(10)
}
  
```

The abstract modeling of QoS 1 works as follows:

- The connection setup between the clients and the broker is the same as that of QoS 0.

- Once the connection is established, the publishing client publishes a message to the broker. The *PUBLISH* packet will contain the client ID (_pid), topic name. After the publish message has been sent, the publisher stores the message.

```
message!PUBLISH,_pid,topic
```

- The subscribing client, on the other hand, sends a *SUBSCRIBE* packet to the broker, specifying its client ID and topic name to which it needs a subscription.
- The broker has stored the message sent by the publisher. Upon getting subscription requests, the broker compares the topic name sent by the subscriber and the topic name created by the publisher. If the name matches (which indicates that both clients are interested in the same topic), the broker publishes the message to all the subscribers (if there are multiple subscribers).

```
publish:
  message!PUBLISH,_pid,topic
```

- After the message is published, the broker sends a *PUBACK*, that is an acknowledgment message stating that the message has been delivered, to the publisher.

```
::message?PUBLISH,packet_id,topic
-> message!PUBACK,packet_id
```

- The publisher then discards the message that it had stored.

If *PUBACK* is lost, the message is retransmitted until the publisher receives an acknowledgment.

3.3.3 QoS 2 (Exactly Once)

QoS 2 is the safest, highest but the slowest level of service. It guarantees that each message is received only once by the receivers. To ensure the message delivery between clients, we use a four-part handshake method (two requests and two response flows) in QoS 2.

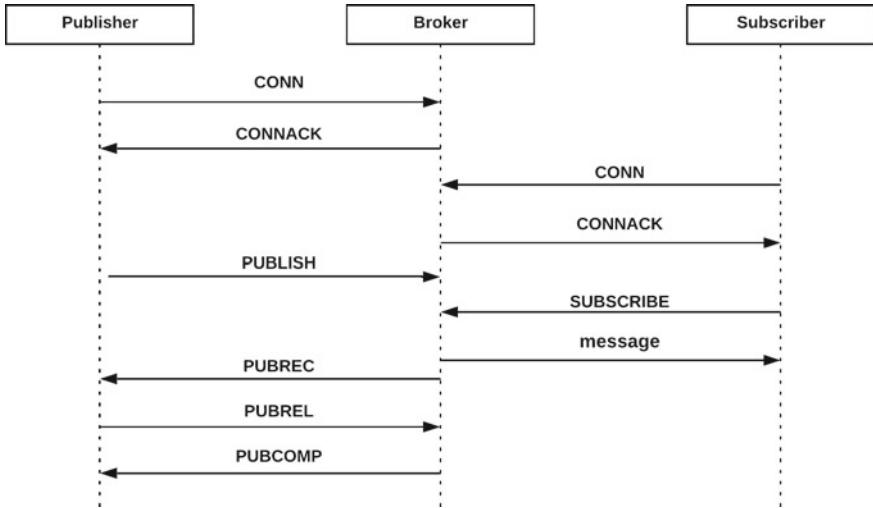


Fig. 4 Sequence diagram of QoS 2

```

init{
    run client(10,20)
    run subscriber(20)
    run subscriber(20)
}
    
```

In QoS 2 (Refer Fig. 4):

- Establishment of connection between the clients and the broker is the same as that of QoS 0 and QoS 1.
- Once the connection establishes, the publisher sends a PUBLISH packet and the subscribing client sends a *SUBSCRIBE* packet to the broker. The publisher will store the message that has been sent.

```

publisher!PUBLISH, client_id,
packet_id, topic_name, message,
qos_type;

subscriber!SUBSCRIBE, client_id,
publisher_id, packet_id, topic_name,
qos_type;
    
```

- The broker then compares both the topic names extracted from the *PUBLISH* and *SUBSCRIBE* packets. If the topic name matches, the message is transmitted to all the subscribers via the broker.

- The broker then sends a *PUBREC* packet to the publisher, which is an acknowledgement to the *PUBLISH* packet. The *PUBLISH* packet is sent again if the sender does not get a *PUBREC* packet from the broker.

channel ! PUBREC, packet_id

- Upon receiving the *PUBREC* packet, the publisher discards the initial *PUBLISH* packet and sends a *PUBREL* packet back to the broker. The publisher stores the *PUBREC* packet.

channel ! PUBREL, packet_id

- After the broker gets the *PUBREL* it discards all stored states and messages. Later the broker sends a *PUBCOMP* packet to the publisher as a response to the *PUBREL* packet.

channel ! PUBCOMP, packet_id

- Finally, the publisher deletes all the other packets and messages.

Similar to QoS 0 and QoS 1, a *timeout* session is created to check if the client has received the *CONNACK* packet. If the waiting time of the client exceeds a value, the client will resend the *CONNECT* packet.

When the message passing is complete in QoS 2, all the clients are sure that the message is delivered. Message retransmissions will take place at a reasonable amount of time if a message is lost along the way.

Once the abstract model of the properties are fed into the SPIN Model checker, we get state-space diagrams as our result. This helps to identify if the clients communicate with each other as per the specified rules.

3.4 Formal Verification

Formal verification is done for the three Quality of Service levels, considering their unique properties. The properties are inserted in the validation model and verified for the message passing interactions between the clients. Verification is done using Linear Time Temporal Logic (LTL) formulae.

Property 1: When a *CONNECT* packet is sent by the clients to the broker, the broker eventually replies by sending a *CONNACK* packet back to the clients.

The connection establishment is a common property of all the three service levels. The LTL identical to this property is as follows:

```
ltl p1 { [] connectionsend -> <>connectionack }
```

connectionsend and *connectionack* are boolean variables. *connectionsend* is set true when the client delivers a *CONNECT* packet to the broker requesting for connection establishment and *connectionack* is set true when the broker eventually sends a *CONNACK* packet back to the client, which indicates that the connection has been successful.

The model checker performs a full state-space search and the property is satisfied.

3.4.1 Properties of QoS 0

The first property of QoS 0 is the connection establishment and receiving the connection acknowledgment.

Property 2: When the publisher sends a message to the subscribing clients, the subscriber may receive the message or not. There is no guarantee of message delivery.

```
ltl p2 {<>publishing->(receivemessage||!receivemessage)}
```

The boolean variables are *publishing*, *receivemessage*, and *!receivemessage*. *publishing* is set true when the publisher sends the *PUBLISH* packet to the broker. We have assumed that the *PUBLISH* packet is eventually sent to the broker without fail, thus, the *publishing* variable is always true. Since there is no guarantee of message delivery, *receivemessage* variable can be true or false.

3.4.2 Properties of QoS 1

The first property of QoS 1 is similar to the first property of QoS 0 (refer Fig. 5), that is, the connection establishment and receiving the connection acknowledgment (Fig. 6).

Property 2: When the publisher publishes a topic, the subscriber will eventually get the topic.

```
ltl p2 { [] topicpublishing -> <> topicsubscribing }
```

The boolean variables *topicpublishing* and *topicsubscribing* is set true then a publishing client creates a topic and sends to the broker through the *PUBLISH* packet and the subscribing client eventually subscribes to the same topic using *SUBSCRIBE* packet.

```

gcc -DMEMLIM=1024 -O2 -DXUSAFE -w -o pan pan.c
./pan -m10000 -a -N p1
Pid: 2683
pan: ltl formula p1

(Spin Version 6.4.9 -- 17 December 2018)
+ Partial Order Reduction

Full statespace search for:
never claim      + (p1)
assertion violations + (if within scope of claim)
acceptance cycles + (fairness disabled)
invalid end states - (disabled by never claim)

State-vector 108 byte, depth reached 0, errors: 0
1 states, stored
0 states, matched
1 transitions (= stored+matched)
0 atomic steps
hash conflicts: 0 (resolved)

Stats on memory usage (in Megabytes):
0.000 equivalent memory usage for states (stored*(State-vector + overhead))
0.283 actual memory usage for states
128.000 memory used for hash table (-w24)
0.534 memory used for DFS stack (-m10000)
128.730 total actual memory usage

```

Fig. 5 Verification result of Property 1 of QoS 0, QoS 1, and QoS 2

Property 3: When the publisher sends the *PUBLISH* packet that includes the topic name and the message, the broker will always send an acknowledgment packet, *PUBACK*, to the publisher, after it sends the message to the subscribing clients.

```
ltl p3 { [] (topicpublishing && receivemessage) -> publishack}
```

topicpublishing, *receivemessage*, and *publishack* are boolean variables. *topicpublishing* is set true when the *PUBLISH* packet is sent by the sender to the broker. *receivemessage* is set true when the broker sends the message to the subscribers and they receive it successfully. Once the message has been received by the subscribers, the broker will send a *PUBACK* packet to the publisher, thus setting *publishack* variable true. In this scenario, we assume that no *PUBACK* packets are lost.

```

gcc -DMEMLIMIT=1024 -O2 -DXUSAFE -DNOCLAIM -w -o pan pan.c
./pan -m10000 -a
Pid: 2065

(Spin Version 6.4.9 -- 17 December 2018)
+ Partial Order Reduction

Full statespace search for:
    never claim      - (not selected)
    assertion violations +
    acceptance cycles + (fairness disabled)
    invalid end states +

State-vector 180 byte, depth reached 142, errors: 0
    12453 states, stored
    12591 states, matched
    25044 transitions (= stored+matched)
    0 atomic steps
hash conflicts:     8 (resolved)

Stats on memory usage (in Megabytes):
    2.470 equivalent memory usage for states (stored*(State-vector + overhead))
    1.847 actual memory usage for states (compression: 74.77%)
        state-vector as stored = 128 byte + 28 byte overhead
    128.000 memory used for hash table (-w24)
    0.534 memory used for DFS stack (-m10000)
    130.292 total actual memory usage

```

Fig. 6 Verification result of Property 2 of QoS 0

3.4.3 Properties of QoS 2

Property 1 is the connection establishment and receiving the connection acknowledgement (refer Fig. 5). The second and the third property of QoS 2 is the same as the properties of QoS 1. Property 2 establishes that when the publisher publishes a topic, any of the clients will eventually subscribe to it (refer Fig. 7). The third property states that after the subscribers receive the message published by the publisher, the broker will send a *PUBREC* packet to the publisher showing that the message has been received by all the subscribing clients (refer Fig. 8).

Property 4: The publisher will always send a *PUBLISH* packet after it receives a *PUBLISH* packet from the broker.

```
ltl p4 {[ ]} sendpublishrec -> <>rcvpublishrel}
```

```

gcc -DMEMLIM=1024 -O2 -DXUSAFE -DNOCLAIM -w -o pan pan.c
./pan -m10000 -a
Pid: 2149

(Spin Version 6.4.9 -- 17 December 2018)
+ Partial Order Reduction

Full statespace search for:
    never claim      - (not selected)
    assertion violations +
    acceptance cycles + (fairness disabled)
    invalid end states +

State-vector 160 byte, depth reached 56, errors: 0
    419 states, stored
    157 states, matched
    576 transitions (= stored+matched)
    0 atomic steps
hash conflicts:      0 (resolved)

Stats on memory usage (in Megabytes):
    0.075 equivalent memory usage for states (stored*(State-vector + overhead))
    0.284 actual memory usage for states
    128.000 memory used for hash table (-w24)
    0.534 memory used for DFS stack (-m10000)
    128.730 total actual memory usage

```

Fig. 7 Verification result of Property 2 of QoS 1

The boolean variable *sendpublishrec* is sent true when the publisher receives a *PUBREC* packet from the broker. *rcvpublishrel* is a boolean variable, which is eventually set true when the publisher sends a *PUBREL* packet to the broker as an acknowledgment to the *PUBREC* packet that it had received (Fig. 9).

Property 5: The broker always send a *PUBCOMP* packet, in response to the *PUBREL* packet sent by the publisher.

```
ltl p5 {[ ]} sendpublishrel-> <>rcvpublishcomp}
```

The boolean variable *sendpublishrel* is sent true when the broker receives the *PUBREL* packet from the publishing client. *rcvpublishcomp* is a boolean variable, that is eventually set true when the broker sends a *PUBCOMP* packet to the publisher as an acknowledgment to the *PUBREL* packet. Figure 10 shows the result after the verification is performed.

```

gcc -DMEMLIM=1024 -O2 -DXUSAFE -DNOCLAIM -w -o pan pan.c
./pan -m10000 -a
Pid: 2280

(Spin Version 6.4.9 -- 17 December 2018)
+ Partial Order Reduction

Full statespace search for:
    never claim      - (not selected)
    assertion violations +
    acceptance cycles + (fairness disabled)
    invalid end states +

State-vector 160 byte, depth reached 56, errors: 0
    419 states, stored
    157 states, matched
    576 transitions (= stored+matched)
    0 atomic steps
hash conflicts:      0 (resolved)

Stats on memory usage (in Megabytes):
    0.075 equivalent memory usage for states (stored*(State-vector + overhead))
    0.284 actual memory usage for states
    128.000 memory used for hash table (-w24)
    0.534 memory used for DFS stack (-m10000)
    128.730 total actual memory usage

```

Fig. 8 Verification result of Property 3 of QoS 1

4 Conclusion and Future Works

A verification model for IoT protocol, Message Queuing Telemetry Transport (MQTT), considering its Quality of Service feature was proposed. The proposed concept was depicted by conducting formal verification of the QoS levels used for communication between different clients. First, an abstract model of the QoS levels was created to analyze the message passing sequence between the publishers, subscribers and the broker. The state-space diagram generated by the SPIN Model Checker gives a detailed representation of the message transactions between the clients. Second, formal verification of the QoS levels is proposed in order to certify its correctness and reliability. The properties of the QoS levels were formulated using linear temporal logic (LTL). The validation model for MQTT was built using PROMELA and verified using the SPIN Model Checker.

```
gcc -DMEMLIM=1024 -O2 -DXUSAFE -DNOCCLAIM -w -o pan pan.c
./pan -m10000 -a
Pid: 2463

(Spin Version 6.4.9 -- 17 December 2018)
+ Partial Order Reduction

Full statespace search for:
    never claim      - (not selected)
    assertion violations +
    acceptance cycles + (fairness disabled)
    invalid end states +

State-vector 156 byte, depth reached 56, errors: 0
    65 states, stored
    2 states, matched
    67 transitions (= stored+matched)
    0 atomic steps
hash conflicts:    0 (resolved)

Stats on memory usage (in Megabytes):
    0.011 equivalent memory usage for states (stored*(State-vector + overhead))
    0.281 actual memory usage for states
    128.000 memory used for hash table (-w24)
    0.534 memory used for DFS stack (-m10000)
    128.730 total actual memory usage
```

Fig. 9 Verification result of Property 4 of QoS 2

As future work, we intend to verify more complex conditions, like ungraceful disconnections, message storing properties, and security-related properties of the QoS levels.

```

gcc -DMEMLIM=1024 -O2 -DXUSAFE -DNOCLAIM -w -o pan pan.c
./pan -m10000 -a
Pid: 2511

(Spin Version 6.4.9 -- 17 December 2018)
+ Partial Order Reduction

Full statespace search for:
    never claim      - (not selected)
    assertion violations +
    acceptance cycles + (fairness disabled)
    invalid end states +

State-vector 156 byte, depth reached 56, errors: 0
    65 states, stored
    2 states, matched
    67 transitions (= stored+matched)
    0 atomic steps
hash conflicts:      0 (resolved)

Stats on memory usage (in Megabytes):
    0.011 equivalent memory usage for states (stored*(State-vector + overhead))
    0.281 actual memory usage for states
    128.000 memory used for hash table (-w24)
    0.534 memory used for DFS stack (-m10000)
    128.730 total actual memory usage

```

Fig. 10 Verification result of Property 5 of QoS 2

References

1. M. Diwan, M. DSouza, A framework for modeling and verifying IoT communication protocols, in *International Symposium on Dependable Software Engineering: Theories, Tools, and Applications* (Springer, 2017), pp. 266–280
2. P. Anudeep, N.K. Prakash, Intelligent passenger information system using IoT for smart cities, in *Smart Innovations in Communication and Computational Sciences* (Springer, 2019), pp. 67–76
3. S.L. Narayan, E. Kavinkartik, E. Prabhu, IoT based food inventory tracking system, in *International Symposium on Signal Processing and Intelligent Recognition Systems* (Springer, 2018), pp. 41–52
4. S. Chouali, A. Boukerche, A. Mostefaoui, Towards a formal analysis of MQTT protocol in the context of communicating vehicles, in *Proceedings of the 15th ACM International Symposium on Mobility Management and Wireless Access* (ACM, 2017), pp. 129–136
5. B. Jayaraman, J.M. Kannimoola, K. Achuthan, Sybil attack detection in vehicular networks, in *Security and Privacy in Internet of Things (IoTs)* (CRC Press, 2016), pp. 55–72
6. N. Mannilthodi, J.M. Kannimoola, Secure IoT: an improbable reality, in *IoTBDS* (2017), pp. 338–343
7. A.R. Tena, L.M. Kristensen, A. Rutle, On Modelling and Validation of the MQTT IoT Protocol for M2M Communication (2018)
8. B. Aziz, A formal model and analysis of an IoT protocol. *Ad Hoc Netw.* **36**, 49–57 (2016)

9. A. Banks, R. Gupta, MQTT Version 3.1. 1. OASIS standard 29, 89 (2014)
10. A.J. Vattakunnel, N.S. Kumar, G.S. Kumar, Modelling and verification of CoAP over routing layer using SPIN model checker. Procedia Comput. Sci. **93**, 299–308 (2016)
11. G.J. Holzmann, *The SPIN Model Checker: Primer and Reference Manual*, vol. 1003 (Addison-Wesley, Reading, 2004)

Prediction of Gene Selection Features Using Improved Multi-objective Spotted Hyena Optimization Algorithm



S. Divya, Eranki L. N. Kiran, Madhu Sudana Rao and Pujitha Vemulapati

Abstract Microarray data analysis is one of the main research areas in the medical research. The Microarray is a dataset which consists of different gene expressions from which most of the features are redundant genes and reducing the classifier accuracy. Finding a minimal subset of features from large gene expression is a challenging task where removing redundant feature but the important feature will not be missed. Many optimization techniques are introduced by the researchers to find a minimal subset of features but it does not provide a feasible solution. In this paper, the RWeka package, which provides an interface of Weka tool functionality to R is used to order the features using select attribute function in Weka. By using those ordered features, a minimal subset of features is selected using SVM classifier with maximum prediction accuracy in the dataset. Obtained minimal subset of features is given as input to the Multi-Objective Spotted Hyena Optimizer algorithm which is driven by the ensemble of SVM classifier by updating the search agents with objective function with an intention to improve the classification accuracy. The proposed method has experimented with seven publicly available microarray datasets such as CNS, colon, leukemia, lymphoma, lung, MLL, and SRBCT, which shows that the proposed methodology gives the high accuracy than all other existing techniques in terms of feature selection and prediction accuracy.

Keywords MOSHO · Microarray data · Gene expression · Feature selection

S. Divya · E. L. N. Kiran (✉) · P. Vemulapati
School of Computing, SASTRA Deemed University, Thanjavur 613402,
Tamil Nadu, India
e-mail: erankikiran@cse.sastra.edu; erankikiran@gmail.com

S. Divya
e-mail: divyasubbu1296@gmail.com

P. Vemulapati
e-mail: pujithavemulapati@gmail.com

M. S. Rao
Department of Mathematics, Amrita Vishwa Vidyapeetham, Coimbatore,
Tamil Nadu, India
e-mail: Madhu031083@gmail.com

1 Introduction

Human body is made up of numerous numbers of cells and each cell is a copy of another cell that should be encoded in DNA. A part or segment of DNA is called as Genes. Human DNA consists of several genes and these genes were used for analyzing the cancer or any other rare type of diseases. Genes are transformed into expressions called as Gene expressions. This type of analysis was taken place in the microarray datasets. Microarray is a hybridized one that takes the particular tissue and labels all the unknown molecular cells. This type of labeling will help us to compare two or more genes and then identify the diseases. Recently, microarray technique is one of the widely used methods for analyzing the gene expression. In microarray data, the accuracy mostly depends on the classification model provided. Experimental results of microarray dataset show better accuracy rate and robustness. This method is suitable for small dimensions or small number of samples. Large dimension datasets can produce noises and fluctuation errors [9]. In this microarray data, the large dataset consist of more number of genes, which contains more redundant genes. Prediction of informative gene from the microarray dataset helps in feature selection. To overcome this problem finding, the informative gene from the microarray dataset feature selection is important. Feature selection is selecting the informative gene from the complex gene expression which explains the data clearly. Feature is the important property for measuring the observations. This selection process overcomes the problem of Curse of Dimensionality and reduces the noise produced in the Over fitting gene expression profile. Feature selection method is suitable for both supervised and unsupervised learning. In supervised learning, the class labels are known and hence it is easy to extract the relevant features from the entire labels. Whereas, in the case of Unsupervised learning, it is quite difficult to extract the features since the class labels are not known. Evolutionary algorithm is applied to microarray data to increase the classification accuracy on the dataset. It is very effective for large dataset and it reduces the computational cost [7]. These algorithms are popular in healthcare analytics as well [1, 11, 12, 14]. Evolutionary algorithm contains several heuristics to solve the optimization task. The general process of Evolutionary Algorithms is random initial population, choose the best individual from the population to generate the next generation, and create the next generation [10].

2 Prior Research Studies

Several research have developed different types of algorithms, which are emerged in few decades for analyzing the gene expression and predicting the accuracy. Numerous approaches in machine learning techniques have been employed by various authors to solve this problem. Few of this recent work have been discussed in this paper. To find the informative gene from the microarray data, Mohamad et.al [10]

presented a hybrid Genetic algorithm. In this hybrid Genetic algorithm, combined SVM classifier is used for prediction accuracy but it is not suitable for multi-class classification problems and the search performance is poor. A hybrid filter and wrapper model feature selection is used for finding the small subset of features from the microarray dataset. In this enhanced binary particle, swarm optimization is used for prediction accuracy. The hybrid model of both filter and wrapper method identifies the most relevant features from the large dataset [4]. SVM and KNN classifiers are used for prediction accuracy but the number of features selected was more and it leads to more number of iterations. In most of the techniques, SVM is used for predicting the accuracy [15]. A LSEFS Linear square support vector machine along with particle swarm optimization is used for feature selection. In this evolutionary feature, selection is also used for selecting the subset of features. SVM is used for predicting the accuracy. Comparing with evolutionary feature selection particle swarm optimization performance is poor. In LSEFS time complexity is high. In this research, single-layer feedforward neural network along with single-value decomposition neural classifier is used for predicting the accuracy [8]. It is not suitable for multi-class classification. Different evolutionary algorithms are used for improving the gene selection and predicting the accuracy. Yu et.al [16] presented an ant colony optimization algorithm to find the selected genes. To enhance the algorithm, fuzzy logic control theory is applied to adjust the parameters, which help to find the appropriate gene with small subset of features it produces high classification accuracy but it requires more number of iterations [3]. It is used to pick the predictive and information gene features from microarray dataset and with selected number of features SVM classifier is used for predicting the accuracy. SVM is used for predicting both binary class and multi-class classification. It proves that ant colony optimization is suitable for high-dimensional, noise, irrelevant, and redundant dataset [6]. Among ant colony optimisation, Naïve Bayes and KNN methods, classification accuracy of naïve bayes and KNN shows better results. However, more advanced ant colony optimization model will reduce the time required for future selection. Bio-Inspired gene selection method (GA, PSO) and accurate wrapper gene selection method are used for feature selection [2]. This feature selection is very effective for large dataset and is capable of searching optimal and near-optimal solutions. It provides high classification accuracy with minimal number of selected genes but the computational cost is high. Sharaf et al. [13] presented a different ranking method for feature selection such as T-test, Fisher score. Cellular Learning Automata-Ant Colony Optimization algorithm (CLA-ACO) is used for selecting the subset of features. For predicting the accuracy SVM, KNN, Naïve Bayes classifiers are used. It also provides high classification accuracy with Fisher Score.

3 Research Approach

The main objective of proposed system is to improve the feature selection from large datasets using SVM classifier to improve the effectiveness of performance.

Feature Selection Using RWeka In this section, we describe the selection of feature subset from large dataset using RWeka. We have applied SVM classifier to measure the predictive accuracy of gene selection parameters. The SVM provides good results comparing with other classifiers. Weka is the tool, which consists of preprocessing, classification, select attributes, and clustering techniques. It consists of three graphical user interfaces, which are Explorer, Experimenter, and Knowledge Flow. From this weka tool, large dataset are applied directly for different algorithms. In this proposed method, we use RWeka package for feature selection. Rweka package interfaces weka functionality to R through set of codes. For installing *RWeka package, RJava, and RWekajars*, package are important from which it provides external jars needed for Rweka. The Rweka package is used for feature ordering using select attribute function in weka. We identified attribute selection measures, namely InfoGain to evaluate the relevance of attribute class and its features in the given dataset.

$$\text{InfoGain(Class, Attribute)} = H(\text{Class}) - H(\text{Class} - \text{Attribute})$$

$$\text{InforGainAttributeEval(formula, data, subset)}$$

- InforGainAttributeEval—R interfaces to Weka attribute Evaluators.
- Formula—The generic function formula and its specific methods provide a way of extracting formulae which have been included in other objects.
- Data—It is used to load specific datasets and it contains the variables in the model.
- Subset—It returns the subset which satisfies the conditions. By using the above formula R interfaces to Weka and it provides the feature ordering of the dataset. From this feature ordering the dataset which has high prediction accuracy will be selected using SVM classifier.

Support Vector Machine SVM is effective classification algorithm, which provides best result when compared with other classifiers. SVM works effectively for data with high-dimensional dataset [17]. At first, SVM works effective for binary class classification only. Later, it was developed and works for multi-class classification. It is well suited for both binary and multi-class classifications. The samples are separated into high-dimensional space where it can be separated by a hyperplane. This function can be carried out by a mechanism called kernel. Kernel is transforming the data into another dimension from which it can be clearly classified between classes of data.

Spotted Hyena Optimization Algorithm Dhiman et.al proposed metaheuristic Spotted Hyena Optimization (SHO) algorithm. The main concept of (SHO) is social behavior of spotted hyena. Hyenas are a dog-like African mammal. The spotted hyenas are skilful hunters and also known as laughing hyenas where their laughing is similar to the human behavior of laughing, and it had spots on their fur reddish brown in color with black spots. Spotted hyenas are arduous, brainy with literally awful character, which have the capability to fight ceaselessly for terrain and food-stuff [5]. In addition to this, female groups are more assertive as compared to male

hyenas and preferred to live in their herds. As male members grow they look for new sects leaving the old one behinds. In a new clan, the male members are the lowest members to get there part of the food. A male member who joined recently to the new sect will always want to join with their new sect family for the long period. Although the female is constantly settled in a permanent place. In distinct, the spotted hyena produces different vibrants to connect with each other over the hunting of their foods. To offer the multi-object edition of SHO first, the basic concepts of SHO is discussed. There are four steps in the SHO algorithm which are listed below.

1. *Encircling Prey*—In this step, we identify the closest solution based on the other search agents inputs while determining the nearest position of the prey. The arithmetic form of the solution provided is as follows:

$$\vec{D}_h = \vec{B} \cdot P_p(i) - P(i) \quad (3)$$

$$\vec{P}(i+1) = P_p(i) - F \cdot D_h \quad (4)$$

where \vec{D}_h represent the accessibility of spotted hyena to the target prey. i indicates the number of sequential rollout. $P_p(i)$ indicate the position of the target vector whereas \vec{P} represents the vectorized position of the selected hyena. The vector \vec{S} and \vec{F} are computed as

$$\vec{S} = 2 \cdot r \vec{d}_1 \quad (5)$$

$$\vec{F} = 2i \cdot r \vec{d}_2 - \vec{t} \quad (6)$$

$$\vec{t} = 5 - (\text{iteration} \times \frac{5}{Max_{\text{iterations}}}) \quad (7)$$

where iterations = 0, 1, 2, 3 ... $Max_{\text{iterations}}$.

2. *Hunting*—In order to predict the target hyena while hunting for the prey and also determine its target space by using the following equation:

$$\vec{D}_h = | \vec{S} \cdot \vec{H}_h - \vec{H}_k | \quad (8)$$

$$\vec{H} = \vec{H}_h - \vec{F} \cdot \vec{H}_h \quad (9)$$

$$\vec{C}_h = \vec{H}_k + \vec{H}_{k+1} + \dots + \vec{H}_{k+N} \quad (10)$$

where \vec{H}_h defines the best optimal hyena selected, H_k represents the reference to other selective hyenas. However, variable N indicates the selective count of most optimal hyena identified which is referred in the equation as follows:

$$N = count_{tos}(H_h, H_{h+1}, H_{h+2}, \dots, (H_{h+o})) \quad (11)$$

where O is a random selective range between $[0.5, 1]$, “ $count_{tos}$ ” represents total number of solutions including candidate solutions, and C_h is a set of most feasible optimal solutions.

3. Attacking—The mathematical representation of prey attack is defined as follows:

$$\vec{P}(i+1) = C_h/N \quad (12)$$

where $\vec{P}(i+1)$ computes the most feasible solution and updates with reference to other search spaces with non feasible solutions.

4. Searching—To search the suitable solution, F is responsible using Eq. (6). Another constituent of proposed SHO algorithm which makes possible for exploration is B. The S vectors contain accidental values which provide the most feasible solution to predict the prey as shown in Eq. (5). Followed by the attack behavior of the hyena using SHO algorithm, assume vector $S > 1$ considering more important than $S < 1$ to project the most feasible solution. To achieve the optimization, SHO algorithm applies randomized selection of solution spaces from the given population. The proposed SHO algorithm can be applied to high dimensional multi-objective problems and solve optimization problem very effectively.

MOSH To upgrade our existing SHO algorithm to support Multi-Objective feature selection, we have applied two mechanisms. The first mechanism is an archive approach, which uses to store non-dominated Pareto optimal solutions and also it sorts the optimal solutions. The second mechanism is a group selection approach in which it selects the adjacent solutions parallel to the location of the prey from archive. These two mechanisms are discussed elaborately in the upcoming sections.

- *Archive*—The best of all non-dominated Pareto optimal solutions are stored in archive. This archive uniformly spread on Pareto front applies solution spaces of uneven distribution. It subsists of two main aspects, namely archive controller and Grid.
- *Archive Controller*—The vital function is to decide whether the solution should be added or not in the archive. The updating rules for archive controller are given below. The current solution is accepted only if the archive is empty. The solution space is eliminated even if one of the archive outputs turns out to be uneven. New solution space is consider into the archive if it is capable of eliminating an existing solution space dominated by the members of the archive. The new solution is stored in the archive if none of the elements contained in the archive show prominence to one another.
- *Grid*—The grid method is used to obtain distributed Pareto fronts. It consists of four linearly separable regions for objective functions. The grid mechanism is mainly used for the computation of each individual from the population. Only those individuals to lie within the grid area are considered for the selection process. The grid method space is constructed by hypercube and it results in uniform distribution.

- *Group Selection Mechanism*—The most challenging issue in multi-objective is comparing the search space with existing search space in archive members. This challenge can be overcome by using group selection mechanism. In this group selection, it chooses least crowded section and it is populated as one of the best solutions in the group of nearby solutions using the roulette wheel mechanism H_k is defined as follows:

$$H_k = \frac{f}{S_k} \quad (13)$$

where f is a constant variable with value greater than 1. S_k represents the number of Pareto optimal solutions to k th segment. This method popularly uses the classical method which defines the contribution of each individual using *roulette wheel* proportion.

4 Results and Discussion

We now present the details of dataset used and the findings of our analysis.

Dataset Used The validation of our proposed method was to test under seven publicly available microarray datasets. This datasets are often used to test the effectiveness of gene selection and classification. We provide the details of the preprocessed datasets used for our study in Table 1.

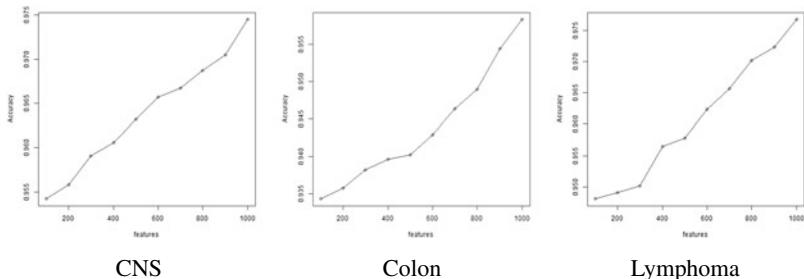


Table 1 Descriptive analysis of preprocessed dataset

Dataset name	Total number of genes	Number of instances	Class
CNS	7129	60	2
COLON	2000	60	2
LEUKEMIA	7129	72	3
LYMPHOMA	4026	62	3
LUNG	12,600	203	4
MLL	12,582	72	3
SRBCT	2308	83	4

The feature selection is applied using RWeka package combining with SVM classifier in which large dataset optimal subset features is extracted and evolutionary algorithm is applied to seven microarray datasets. The execution of the algorithm is decided by surveying the accuracy from the optimal subset features. The above results are shows that the accuracy measure of six datasets. The results of presented algorithm produce higher accuracy with small number of features.

5 Conclusions

The feature selection is applied using R language and SVM of WEKA data mining tool is implemented by using RWeka package. Finding the optimal subset of feature using RWeka package from which info gain-based feature ordering is performed and features with high classification accuracy is selected as optimal subset feature using SVM as classifier. Then, the optimal subset feature is given as input to the Multi-objective Spotted Hyena Optimizer, which is driven by the ensemble of SVM classifier from which it updates search agents of spotted hyenas. The SVM classifier is used for predicting the accuracy of the optimal subset feature. The proposed approach is tested on seven microarray datasets, which shows better performance when compared with other existing approaches. In this, leukemia dataset obtains 100% accuracy. The improvement in the prediction accuracy ranging from 4% to 5% increases when compared with proposed approach.

References

1. K. Akyol, Ü. Atila, A study on performance improvement of heart disease prediction by attribute selection methods. *Akademik Platform Mühendislik ve Fen Bilimleri Dergisi* **7**(2), 174–179
2. H.M. Alshamlan, G.H. Badr, Y.A. Alohal, The performance of bio-inspired evolutionary gene selection methods for cancer classification using microarray dataset. *Int. J. Biosci. Biochem. Bioinform.* **4**(3), 166 (2014)
3. H.M. Alshamlan, G.H. Badr, Y.A. Alohal, Abc-svm: artificial bee colony and svm method for microarray gene selection and multi class cancer classification. *Int. J. Mach. Learn. Comput.* **6**(3), 184 (2016)
4. L.Y. Chuang, C.H. Ke, C.H. Yang, C.H.: A hybrid both filter and wrapper feature selection method for microarray classification (2016), [arXiv:1612.08669](https://arxiv.org/abs/1612.08669)
5. G. Dhiman, V. Kumar, Spotted hyena optimizer: a novel bio-inspired based metaheuristic technique for engineering applications. *Adv. Eng. Softw.* **114**, 48–70 (2017)
6. T.M. Fahrudin, I. Syarif, A.R. Barakbah, Ant colony algorithm for feature selection on microarray datasets, in *2016 International Electronics Symposium (IES)* (IEEE, 2016), pp. 351–356
7. F. Fernández-Navarro, C. Hervás-Martínez, R. Ruiz, J.C. Riquelme, Evolutionary generalized radial basis function neural networks for improving prediction accuracy in gene classification using feature selection. *Appl. Soft Comput.* **12**(6), 1787–1800 (2012)
8. H.T. Huynh, J.J. Kim, Y. Won, Classification study on dna microarray with feedforward neural network trained by singular value decomposition. *Int. J. Bio-Sci. Bio-Technol.* **1**(1), 17–24 (2009)

9. T. Juliusdottir, E. Keedwell, D. Corne, A. Narayanan, Two-phase EA/k-NN for feature selection and classification in cancer microarray datasets, in *2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology* (IEEE, 2005), pp. 1–8
10. M.S. Mohamad, S. Deris, R.M. Illias, A hybrid of genetic algorithm and support vector machine for features selection and classification of gene expression microarray. *Int. J. Comput. Intell. Appl.* **5**(01), 91–107 (2005)
11. M.R. Nalluri, D.S. Roy et al., Hybrid disease diagnosis using multiobjective optimization with evolutionary parameter optimization. *J. Healthc. Eng.* **2017** (2017)
12. N.M. Rao, K. Kannan, X.Z. Gao, D.S. Roy, Novel classifiers for intelligent disease diagnosis with multi-objective parameter evolution. *Comput. Electr. Eng.* **67**, 483–496 (2018)
13. F.V. Sharbaf, S. Mosafer, M.H. Moattar, A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization. *Genomics* **107**(6), 231–238 (2016)
14. T.S. Sujana, N.M.S. Rao, R.S. Reddy, An efficient feature selection using parallel cuckoo search and naïve bayes classifier, in *2017 International Conference on Networks & Advances in Computational Technologies (NetACT)* (IEEE, 2017), pp. 167–172
15. E.K. Tang, P.N. Suganthan, X. Yao, Feature selection for microarray data using least squares svm and particle swarm optimization, in *2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology* (IEEE, 2005), pp. 1–8
16. H. Yu, S. Hong, X. Yang, J. Ni, Y. Dan, B. Qin, Recognition of multiple imbalanced cancer types based on DNA microarray data using ensemble classifiers. *BioMed Res. Int.* **2013** (2013)
17. X. Zhou, D.P. Tuck, MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on dna microarray data. *Bioinformatics* **23**(9), 1106–1114 (2007)

A Compressive Family Based Efficient Trust Routing Protocol (C-FETRP) for Maximizing the Lifetime of WSN



Nandoori Srikanth and Muktyala Siva Ganga Prasad

Abstract WSN is deployed for the dissemination of various sensor nodes in a fixed topology to sense the environment with limited resources, and to communicate the sensed data with the base station through cluster head. WSN is one of the dynamic networks, which can perform several dynamic functions like change in cluster head, avoid redundant messages, resource reservation mechanism and resource cancellation mechanism. One of the major problems in the deployment of WSN is ‘terrain structure’. Due to irregular terrain structure, deployment of sensor nodes is random in nature and due to this random deployment; nodes are not properly organized in a distributed way. A Compressive Family-based Efficient Trust Routing Protocol is proposed by dividing the network into various clusters and then split up clusters into sub-clusters. Each sub-cluster is again separated into various families and each family is allocated with a family head. The performance of proposed approach is compared with similar protocols (GEED-M, EETRP, and FERP) developed for specific terrain structures like plateaus and military areas. The Compressive Family-based Efficient Trust Routing Protocol enhances the shelf life of the network by 69%, and reduces the energy consumption of the network by 30%.

Keywords WSN · Data aggregation · Trust node · Malicious node · Built in self-test

1 Introduction

WSN has been remodelling the human lifestyle frequently for the past few decades; it reforms various fields like medical, industrial, habitat monitoring and traffic controlling. At the initial stages of WSN, there is a limited software development, lack of network support, dependency on traditional methods from industries, and medical

N. Srikanth (✉) · M. Siva Ganga Prasad

Department of ECE, Koneru Lakshmiah Educational Foundation, Vaddeswaram, India

e-mail: srilovesnature@gmail.com

M. Siva Ganga Prasad

e-mail: msivagangaprasad@kluniversity.in

© Springer Nature Singapore Pte Ltd. 2020

L. C. Jain et al. (eds.), *Data Communication and Networks*,

Advances in Intelligent Systems and Computing 1049,

https://doi.org/10.1007/978-981-15-0132-6_6

fields. The development rate of WSN in recent years has been very high, causes improvisations in several disciplines and invokes various new methods to achieve optimized results [1]. Authors proposed various protocols to overcome the limitations of network like increasing the network lifetime, reducing the energy consumption, avoiding malicious attacks and efficient utilization of resources. One of the major applications of WSN is monitoring the military fields. There are so many clustering, routing, and cluster-based routings proposed to make the network more energy efficient [2, 3]. Cluster-based routing techniques give the best results in WSN compared with the existing protocols. Some researchers have come out with mobile data collectors among clusters for data collection; this work engenders the network more energy efficient, and gives the best results in military, plateaus and non-uniform terrain structures. The sensor node task includes sensing, processing, computing and communicating. Sensor nodes work under the principle of sleep scheduling, where sensor nodes can go to sleep state automatically once they finish their sensing round until they get a notification from mobile data collector to send data. The mobile data collector has complete knowledge about each sensor node's position and its residual energies in its assigned sub-cluster, hence mobile data collector can act as a sub-cluster head for particular sub-cluster [4]. The mobile data collector gathers data from each sensor node in its assigned sub-cluster aggregates the data and forwards to the cluster head. Here, mobile data collector can function as a gateway node to sub-cluster nodes. Some highly energized nodes are called trusted nodes [5], they send their sensed data and compare it with neighbour's and then forward to mobile data collectors in a specific time interval allotted by the network.

B. A. Mohan, H. Saroja Devi, proposed an efficient hybrid data collection algorithm [6], for data collection from multiple mobile nodes. In this technique, the cluster head will be elected by the base station for the first two rounds using centralized algorithm, after that CH selection is based on previous cluster heads selection in a distributed way. Here, a mobile node is introduced between CH, and Base Station. For this type of applications, Mobile nodes are assigned with unlimited resources to increase lifetime of Network.

J. Luo and J.-P. Hubaux, presented a energy efficient and conserving routing protocol [7] for the purpose of improving lifetime, by managing the concentration of data traffic at small number of base stations. In WSN, sensors which are nearer to BS have to relay high amount of data traffic, then those nodes batteries ends up quickly. To overcome this problem, the BS should be a mobile, then automatically sensors nearer to BS changes timely, and No more data traffic burden would be on the same nodes.

Atakli et al. developed a scheme based on weighted-trust estimation in order to detect and isolate the compromised nodes in hierarchical clustered WSN structure. In this scheme, they select some nodes as Forwarding Nodes to give a trust values for all of the cluster nodes. Afterwards, they decrease the node's trust level for all nodes that sent malicious information. Tolba et al. [8] proposed an energy efficient algorithm, for mobile WSNs, It is a distributed clustering algorithm, and it is named as ALM. This algorithm enhances network lifetime, and it also improves the stability, and network connectivity.

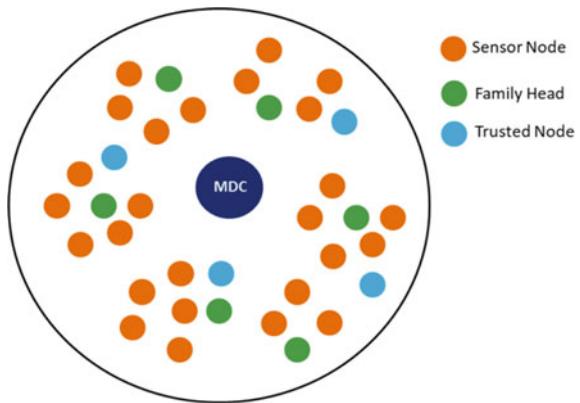
Gong et al. presented a routing protocol for the purpose of energy efficiency and security in WSNs, named, Secure and Energy Aware Routing Protocol (ETARP) [9]. The main contribution point in ETARP is route discovering and selection based on both the maximum utility concept. ETARP scheme takes into consideration the energy efficiency and the trustworthiness in routing protocol, which may sustain more complexity and overhead compared to AODV routing protocol. The Cluster based routing protocols (CBR-MOBILE) is proposed to face the challenges of packet loss and energy consumption in Hybrid networks like some sensor nodes are fixed, and others are mobile. It is traffic adaptive protocol that assigns timeslots of mobile nodes which are moves out of cluster can be reassign to the incoming mobile nodes into that cluster. Based on receiving signal strength, data is transmitted to the cluster head [10]. A secure mobile data collector is introduced in clusters to collect the data from cluster head, and forward to the base station. Authors proposed and analysed three protocols for secure data collection, and it follows tree-based connection management among sensor nodes [11].

S. Deng, J. Li, L. Shen proposed a Mobility based clustering protocol for wireless sensor networks, with mobile nodes. Based on its residual energy and mobility of the sensor node, it can decide itself as a cluster head. Based on connection time estimation, the sensor node aims at link stability, which is connected between sensor node and cluster head. Each sensor node can send its data in assigned time slot in an ascending order (TDMA). During mobility condition; sensor node sends a joining request message to the new cluster head, about its joining when it lost its connection with previous cluster head [12].

2 Proposed Work

A huge number of clustering, routing protocols are proposed and commissioned which are expected to make the wireless sensor network more energy efficient. Non-uniform terrain structures suffer from several limitations right from deployment to communication. To figure out these problems, cluster-based routing protocols have been proposed and they gave good results in comparison with traditional and existing approaches. In present work, A Compressive Family-based Efficient Trust Routing Protocol is proposed to make the network more energy efficient. This protocol is a hybrid routing protocol, achieved by combining the characteristics of Family-based Efficient Routing Protocol, and Energy Efficient Trust node-based Routing Protocol. The structure of sub cluster in C-FETRP protocol is shown in Fig. 1, and the proposed work performance analysis proves that C-FETRP protocol is more skillful than other proposed works, and it shows enhanced results in Packet delivery ratio, Lifetime improvement and Energy efficiency.

Fig. 1 Structure of sub cluster in C-FETRP protocol



2.1 Contributions of the Work

The main contributions of the work are dividing the network into various clusters and then dividing into sub-clusters and allocate each cluster with a cluster head. Apart from this conservative work, the proposed approach has the following contributions:

- Each sub-cluster is assigned with a mobile node for data collection, which is christened as Mobile Data Collector.
- Each sub-cluster is again divided into various groups like families, and each family elects a family head based on its residual energy [13].
- The family head collects data among family members and forwards it to the mobile data collector (MDC).
- Some highly energized nodes are separately allocated as trusted nodes for transmitting data to the mobile data collectors in Trust rounds.

Total Number of nodes	75
Packet size	512 bytes
Rx power	0.075
Compression ratio	$((3.14/4) * 0.075 + 65)/75 = 86\%$
Data rate before compression	$65 * 512 = 33280$ bytes
Data rate after compression	28620.8 bytes

2.2 Compressive Family Based Efficient Trust Routing Protocol

In proposed C-FETRP protocol, network is partitioned into clusters, and assigned a cluster head based on residual energy, and which is at the nearest distance to the host. The cluster is again divided into group of sub-clusters, and each sub-cluster is

divided into various groups called families. Each family elects its family head based on their residual energy, and the elected family head collects data from its family members and forwards it to mobile data collector. The mobile data collector gathers data from family heads instead of collecting the data from all family members [6, 7]. Due to this, the mobile data collector's, and sensor nodes energy consumption is reduced to a great extent. On the other hand some high energized nodes in every family are elected as trusted nodes to send sensed data to mobile data collector in even number of rounds.

The network divides data collection rounds into two categories; Un-trusted round (Odd Round) and Trusted round (Even Round). In Un-trusted round, MDC moves around sub-cluster, gathers data from family heads and forwards to the cluster head after data aggregation. In Trusted round, MDC will remain stationary, whereas trusted nodes will sense, compare sensed data with neighbours and broadcast the processed data to MDC directly. This alternative data gathering approach repeats for entire data collection and improves the network lifetime to a great extent and this will be discussed in results section. Based on application number of trust rounds, count can be increased per one Un-trusted round.

The Data Compression technique is used to reduce the volume of information to be stored into storages or to reduce the communication bandwidth required for its transmission over the networks. Here, we are using Text-based Compression technique. Data compression, source coding, or bit-rate reduction involves encoding information using fewer bits than the original representation. Compression can be either with loss or lossless. Lossless compression reduces bits by identifying and eliminating statistical redundancy.

2.3 Energy Consumption Model

For transmitting an m-bit message over a distance 'n' is

$$E(\text{Trans})_{m,n} = \begin{cases} m \times E(\text{elec}) + n \times \alpha(fs) \times n^2, & n < n(0) \\ m \times E(\text{elec}) + n \times \alpha(mp) \times n^2, & n \geq n(0) \end{cases} \quad (1)$$

$$\text{For receiving } m\text{-bit message is } E(\text{rec})_m = m \times E(\text{elec}) \quad (2)$$

To aggregate K messages with length m-bit is denoted by

$$E(\text{agg})_{k,m} = k \times m \times E(\text{da}) \quad (3)$$

where $E(\text{da})$ is the energy dissipated per bit to aggregate message signal.

$E_{(\text{elec})}$ is the energy consumed by the sensor node for a bit of data transmission. The amount of energy utilized in current round can be expressed as

$$\text{Residual energy} = RE + S(i) * E \quad (4)$$

Average residual energy (ARE) can be calculated by using the formula

$$\text{ARE}(\text{Round} + 1) = \frac{\text{RE}(\text{Round} + 1)}{2} \quad (5)$$

The following equation gives total energy consumption (TEC) on each round.

$$\text{TEC}(\text{Round} + 1) = E_0 * n - \text{RE}(\text{Round} + 1) \quad (6)$$

In case are ‘n’ layers available in the network, the average energy consumption of node can be defined as

$$\text{AEC}(\text{Round} + 1) = \frac{\text{TEC}(\text{Round} + 1)}{n} \quad (7)$$

The AEC is calculated with respect to the total energy consumption. TEC consists of the average of all transmitted energy, received energy, idle energy and sleep mode energy. The result shows the total dead and alive nodes present in the system.

2.4 Energy Consumption Model for Proposed Algorithm

The energy consumption of mobile data collector to send R bits of data to the CH, when it is located at centre of sensing region

$$E_{(\text{MDC})} = m \times E_{(\text{elec})} + m \times E_{(\text{s})} \times r_h^2 \quad (8)$$

where $E_{(\text{MDC})}$ is the energy consumed by mobile data collector node, and r_h is the average distance between mobile data collector and cluster head.

$$r_h^2 = \frac{L^2}{2\pi K} \quad (9)$$

RE is the residual energy consumed by the cluster head for K messages of data transmission in particular round with a data rate ‘m’, and it is expressed as

$$\text{RE} = \left(\frac{T}{K} - 1 \right) \times m \times E_{(\text{elec})} + \frac{T}{K} \times m \times E_{(\text{d})} + m \times E_{(\text{elec})} + \alpha(f_s) \times r_d \quad (10)$$

where T is the number of nodes equally dispersed over the square area $L \times L$.

$E_{(\text{d})}$ is the energy consumed per bit report to the base station, and r_d is the distance between cluster head to the base station.

3 Performance Analysis

The performance of C-FETRP protocol is designed and developed by using network simulator-2 (ns-2), and it is compared with existing protocols which are previously proposed. The proposed protocol C-FETRP gives the best results compared with Green COMP based Energy Efficient Data Aggregation algorithm (GEED-M), Energy Efficient Trust node based routing protocol (EETRP) and the Family based Energy Efficient Routing Protocol (FERP). These three protocols were chosen as the objective of those protocols is same as that of the proposed protocol. C-FETRP performance is evaluated under the following metrics: (i) Energy efficiency, (ii) Energy Consumption, (iii) Throughput and (iv) Network lifetime. From the perspective of implementation, sensor nodes 61, 62, 63, 64 are mobile data collectors, and sensor nodes 21, 14, 55 are trusted nodes. In un-trusted round, MDC moves around sub-cluster, and collects data from family heads, in trusted rounds mobile data collectors remain stationary and trusted nodes send sensed data to MDCs directly. Due to this alternative rounds of data collection, energy consumed by MDCs and sensor nodes can be reduced. Depending on application, we can increase trusted rounds iterations per cycle (Table 1 and Fig. 2).

Network Lifetime

Network lifetime depends on ‘the number of rounds that sensor nodes can withstand with minimum residual energy’. Efficient utilization of energy resources is the key parameter to decide the network lifetime. Increase of network lifetime causes improvement in throughput, energy efficiency and QOS, etc. The proposed C-FETRP

Table 1 Simulation parameters

Parameter	Values
Simulation period	100 ms
Coverage area	1320 * 1032
No. of nodes	75
No. of sink node	1
No. of mobile node	4
No. of sub cluster	4
No. of cluster head	1
No. of trusted nodes	8
Traffic type	CBR
Agent type	UDP
Routing protocol	AODV
Initial power	100 J
Transmission power	1 J
Receiving power	1 J
Queue type	Drop-tail

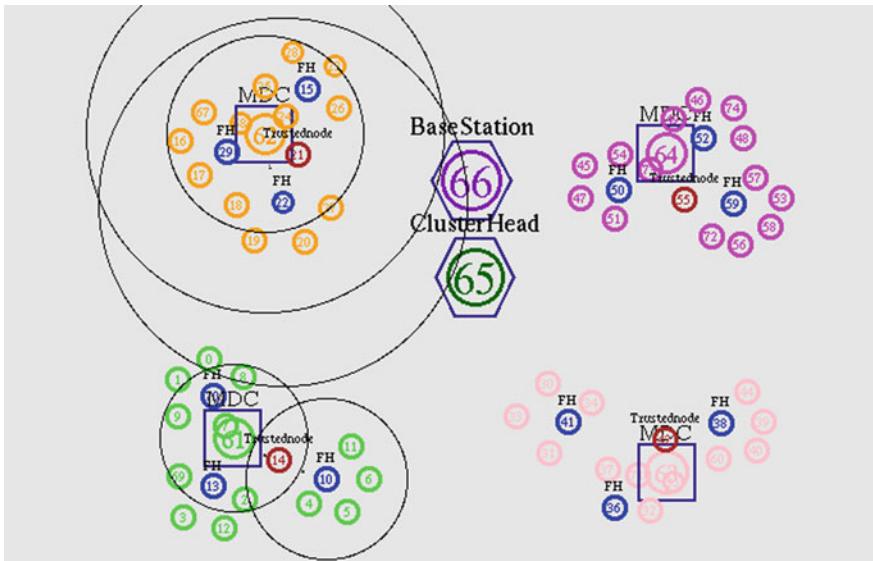


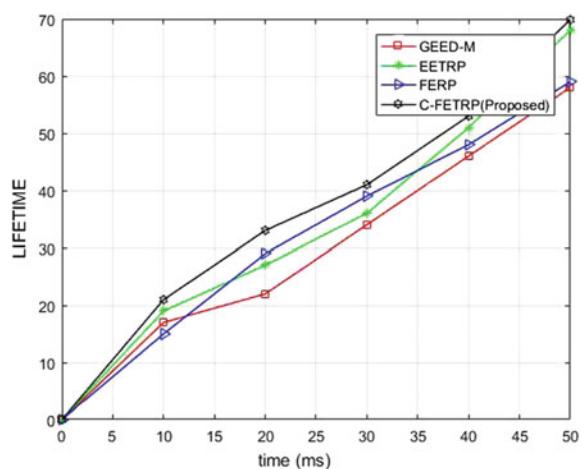
Fig. 2 Implementation of C-FETRP protocol

protocol improves network lifetime up to 69% compared with previously proposed works GEED-M 56%, FERP 59%, and EETRP 66% as shown in Fig. 3.

Energy Consumption

The Energy consumed by the sensor nodes, transceiver, processor, and memory unit leads to energy consumption of the network. The total energy consumed by all nodes can be expressed as

Fig. 3 Network lifetime



$$\text{Energy Consumption} = \sum_{i=1}^n E_i$$

Figure 4 shows energy consumption comparison graphs of various protocols. The energy consumption of C-FETRP protocol is 30% and other protocols GEED-M 44%, FERP 40%, and EETRP 32%.

Network Throughput

The network throughput depends on the amount of packets forwarded by non-base station and amount of packets received by base station. The throughput of C-FETRP protocol is 1426 kbps, and other protocols GEED-M 886 kbps, FERP 1526 kbps, and EETRP 1153 kbps (Figs. 5 and 6)

Fig. 4 Energy consumption

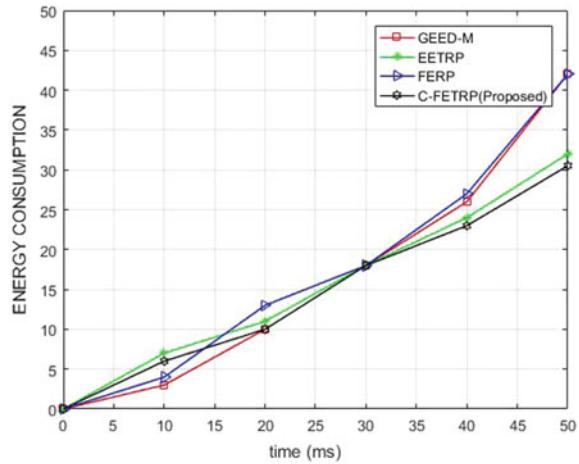


Fig. 5 Throughput

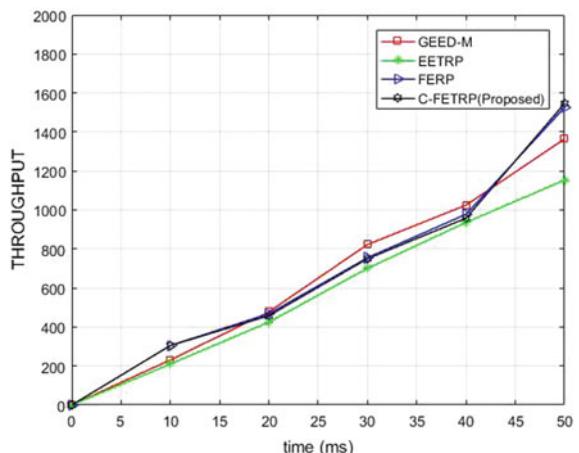
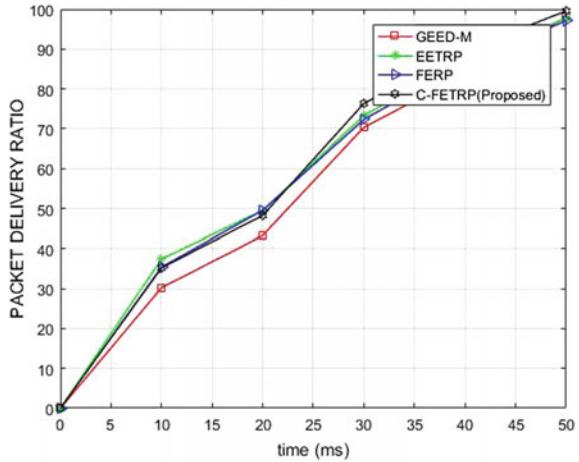


Fig. 6 Packet delivery ratio

Packet Delivery Ratio

It is the ratio of packets generated at the source end to the packets received at the sink in a network. The figure shows comparison of C-FETRP protocol with other proposed protocols in packet delivery. Packet loss causes energy wastage, regeneration of packets at source end, less throughput, and reduced QOS. The proposed protocol shows a linear improvement in PDR compared with other protocols. The packet delivery ratio evaluated in proposed protocol C-FETRP is 99.57% and other protocols GEED-M 97.4%, FERP 97.14%, and EETRP 97.77% (Table 2).

Table 2 Comparison between existing protocols with proposed protocol

Parameter	GEED-M [14]	FERP [15]	EETRP [16]	C-FETRP (proposed)
Packet delivery rate (%)	97.419	97.12	97.77	99.5753
Control overhead	969 packets	952 packets	973 packets	928 packets
Energy consumption (%)	44	42	32	30.1498
Energy efficiency (%)	54	56	68	68.6
Throughput (kbps)	886	1526	1153	1426.65
Loss	207 packets	164 packets	154 packets	107 packets
Lifetime	56%	59% (1400 rounds)	66%	69.8502%

4 Conclusion

Mobile data collector-based routing protocols gives better results compared with conventional and traditional protocols in uneven terrain structures like military areas and plateaus, and where multi-hop communication is complex. The proposed C-FETRP protocol is a hybrid routing protocol which is obtained by combining the characteristics of EETRP and FERP protocols. This protocol is designed and developed by using ns-2 simulator and graphs are generated by using mat lab software for better visibility of experimental values. C-FETRP protocol is compared with previously proposed works, GEED-M, EETRP, and FERP protocols and it gives better results compared to all these protocols. The performance of this routing protocol is assessed based on Energy consumption, Throughput, Lifetime, Packet Delivery Ratio and Energy efficiency. Most of the mobile data collector's energy is saved due to family-based efficient routing and sensor nodes energy is saved due to trust node based routing. By combining these two routing methods, better performance of the network is achieved and progressive improvements in results are obtained.

References

1. L.F. Akyildiz, T. Melodia, K.R. Chowdhury, A survey on wireless multimedia sensor networks. *Comput. Netw.* **51**(4), 921–960 (2007) (Elsevier)
2. O. DurmazIncel, A. Ghosh, B. Krishnamachari, K. Chintalapudi, Fast data collection in tree-based wireless sensor networks. *IEEE Trans. Mob. Comput.* **11**(1), 86–99 (2012)
3. V. Mhatre, C. Rosenberg, Design guidelines for wireless sensor networks communication: clustering and aggregation. *Elsevier Ad Hoc Netw. J.* **2**(1), 45–63 (2004)
4. P.H. Huang, S.S. Sun, W. Liao, GreenCoMP: energy-aware cooperation for green cellular networks. *IEEE Trans. Mob. Comput.* **16**(1), 143–157 (2017)
5. S.M. Sajjad, S.H. Bouk, M. Yousaf, Neighbor node trust based intrusion detection system for WSN. *Procedia Comput. Sci.* **63**, 183–188 (2015)
6. B.A. Mohan, H. Saroja Devi, A hybrid approach for data collection using multiple mobile nodes in WSN (HADMMN), in *Proceedings of IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, vol. 5, (2016), pp. 736–739
7. J. Luo, J.-P. Hubaux, Joint mobility and routing for lifetime elongation in wireless sensor networks, in *Proceedings of IEEE INFOCOM*, vol. 3 (2005), pp. 1735–1746
8. F.D. Tolba, W. Ajib, A. Obaid, Distributed clustering algorithm for mobile wireless sensors networks, in *Proceedings of the 12th IEEE SENSORS 2013, Conference* (2013), pp. 1–4
9. P. Gong, T.M. Chen, Q. Xu, ETARP: energy efficient trust—aware routing protocol for wireless sensor networks. *J. Sens.* (Article ID 469793), 10 (2015)
10. S.A.B. Awwad, C.K. Ng, N.K. Noordin, M.F.A. Rasid, Cluster based routing protocol for mobile nodes in wireless sensor network. *Wirel. Pers. Commun.* **61**(2), 251–281 (2010)
11. A.S. Poornima, B.B. Amberker, Secure data collection using mobile data collector in clustered wireless sensor networks. *IET Wirel. Sens. Syst.* **1**, 85–95 (2011)
12. S. Deng, J. Li, L. Shen, Mobility-based clustering protocol for wireless sensor networks with mobile nodes. *IET Wirel. Sens. Syst.* **1**(1), 39–47 (2011)
13. R. Ferdous, V. Muthukumarasamy, A. Sattar, A node-based trust management scheme for mobile ad-hoc networks, in *Proceedings of the 4th International Conference on Network and System Security (NSS)* (2010), pp. 275–280

14. N. Srikanth, M.S. Prasad, Green comp based energy efficient data aggregation algorithm with malicious node identification (geed-m) for lifetime improvement in WSN. COMPUSOFT, Int. J. Adv. Compt. Technol. **8**(4), 3117–3125 (2019)
15. N. Srikanth, M.S. Prasad, Family based efficient routing protocol for lifetime improvement in heterogeneous WSN. J. Adv. Res. Dyn. Control Syst. **11**(6) (2019)
16. N. Srikanth, M.S. Prasad, Energy efficient trust node based routing protocol (EETRP) to maximize the lifetime of wireless sensor networks in Plateaus. Int. J. Online Eng. **15**(6), 113–130 (2019)

An Adaptive Genetic Co-relation Node Optimization Routing for Wireless Sensor Network



Nandoori Srikanth and Muktyala Siva Ganga Prasad

Abstract Wireless sensor network is designed with low energy, and limited data rates. In wireless sensor networks, the sensors are designed with limited energy rates and bandwidth rates. Maximizing the network lifetime is a key aspect in traditional Wireless communication to maximize the data rate in typical environments. The clustering is an effective topology control approach to organize efficient communication in traditional sensor network models. However, the hierarchical-based clustering approach consumes more energy rates for large-scale networks for data distribution and data gathering process, the selection of efficient cluster and cluster heads (CH) play an import role to achieve the goal. In this paper, we proposed an Adaptive Genetic Co-relation Node Optimization for selecting an optimal number of clusters with cluster heads based on the node status or fitness level. Using the tradition Genetic Algorithm, we achieved the Cluster head selection and the co-relation approach identifies the optimal clusters heads in a network for data distribution. Cluster head election is an important parameter, which leads to energy minimization, and it is implemented by Genetic Algorithm. Appropriate GAs operators such as reproduction, crossover and mutation are developed and tested.

Keywords WSN · GA · Adaptive genetic co-relation node optimization

1 Introduction

Wireless sensor network (WSN) is a self-organized network system with low amount of resources and constitutes of tiny sensors communicate to a remote base station [1]. Nowadays, WSNs are widely used as an effective communication interface medium to interact with physical world to exchange global information. In addition, WSN

N. Srikanth (✉) · M. Siva Ganga Prasad

Department of ECE, Koneru Lakshmi Educational Foundation, Vaddeswaram, Guntur, India
e-mail: srilovesnature@gmail.com

M. Siva Ganga Prasad
e-mail: msivagangaprasad@kluniversity.in

consist of spatially distributed autonomous sensors to cooperatively monitor physical or environmental conditions. Broadcasting across autonomous sensors produces more communication issues due to the lack of resources such as energy, bandwidth, and memory. The recent advances in the microelectromechanicalsystems (MEMS) Technology produced low-cost sensors, as a result, WSNs have paid more attention to different industrial applications [2].

For past few years, an intensive research was conducted to address the problems during data gathering and processing among group of sensors and to address the potential of collaboration among sensors. However, sensor nodes are constrained nodes and organizing large amount of communication services is a problem due to the lack of energy resources and bandwidth. However, the sensors are powered by low-cost irreplaceable batteries which makes for an interesting research to design a new energy-efficient protocol in an unattended hostile environment. Cluster-based protocols are one of the well-accepted protocols and organized the sensors effectively in the network [3]. In this clustering process, the network is divided into different zones, each zone represent as a cluster, each cluster consists of set of sensor nodes and cluster head (CH), the set of sensor nodes are represented as cluster member, the cluster members in each cluster exchange a data with cluster head (CH). The CH distributes the collected data to corresponding destination point. The overall data gathering and data distribution process needs more attention to improve the data distribution rate in typical environments. In order to organize an effective or efficient communication services, there were various cluster-based routing models were designed, i.e. LEACH [4], PEGASIS [5], TEEN [6], and APTEEN [7]. The main limitations of these protocols have identifying optimal clusters and optimal cluster heads (CH) for large scale network due to the exponential variation computational complexity. However the energy-efficient based and topology-based routing protocols address the node fitness issues, an inappropriate cluster and CH selection process increases communication overhead.

We introduce an Adaptive Genetic Co-relation Node Optimization for identifying an optimal cluster and optimal cluster head (CH). The adaptive energy rate allocation scheme identifies the optimal energy of each node, and the optimal nodes are assigned to genetic algorithm to identify the node co-relation. The genetic algorithm computes the node fitness and clustering fitness based on the node characteristics such as energy, distance to sink node, density, and fairness in different stages which is described in Sect. 4. The Genetic Co-relation Node Optimization Routing (GCNO) approach optimize the optimal routing based on the node fitness and fairness level in each cluster and corresponding cluster heads (CHs).

1.1 Contributions of the Paper

The following contribution were designed in this paper:

- We design a cluster-based wireless sensor network model by employing traditional adaptive energy rate allocation scheme.
- Discover an optimal clusters and cluster heads using Adaptive Genetic Co-relation Node Optimization scheme.
- Design a Genetic Co-relation Node Optimization Routing (GCNO) protocol for processing optimal routing.

The paper organizes the following sections, the Sect. 2 describes of related work, which describes the research gaps of various cluster-based routing protocols, Sect. 3 describes the network model and adaptive energy rate model. The Sect. 4 presents the genetic algorithm for optimization of clusters and cluster heads. Section 5 presents the Genetic Co-relation Node Optimization Routing protocol for route optimization. Section 6 presents the experimental student and result discussion (Fig. 1).

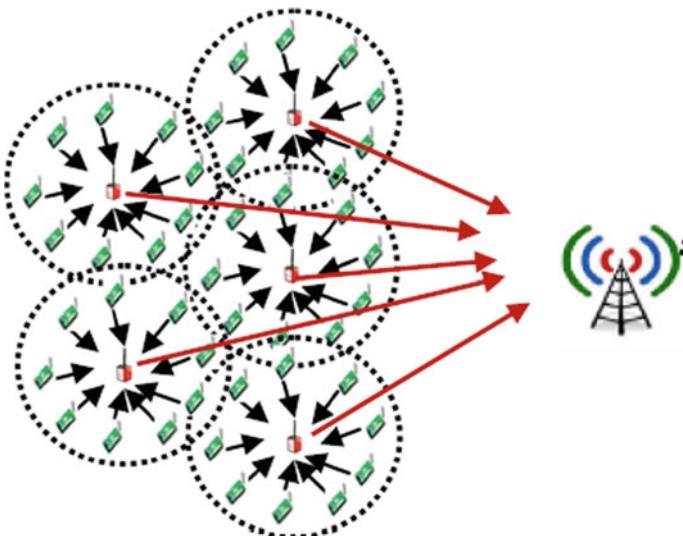


Fig. 1 Cluster-based WSN

2 Related Work

Author	Title	Research methodology	Research gap analysis
Sai Wang, Thu L. N. Nguyen, and Yoan Shin, Senior Member, IEEE (2018) [8]	Energy-Efficient Clustering Algorithm for Magnetic Induction-Based Underwater Wireless Sensor Networks	This paper presents the clustering techniques for MI-based UWSNs. The sensor nodes are configured with the Poisson distribution. To obtain a clustering rules by controlling the energy consumption, based upon the idea of the conventional HENPC (High energy node priority clustering) The proposed method is a dynamic protocol that is CH (cluster head) selection depends upon the remaining energy. From the side of saving energy, multi-hop data and nodes with high remaining energy preferred to be selected as CH(cluster head)s, can efficiently maintain the energy consumption for whole network	The clustering process is based on energy model, which is a tedious process, every time node energy state get varies which impacts on organizing clusters and increases network lifetime
Yi Zhou, Shubhhi Taneja, Chaowei Zhang, Xiao Qin, Senior Member, IEEE (2018) [9]	GreenDB: Energy-Efficient Prefetching and Caching in Database Clusters	This paper described about a parallel database system called Green DB, which is a energy-efficient system for clusters. The main feature of Green DB is a caching mechanism, which receives node or route information from passive nodes into active nodes. Green DB organizes an information table to maintain the nodes state information. This protocol designs a congestion-free route model to save energy and provides optimal route	Here the major limitation is, for construction of route data information by considering transnational database systems. This data saving process frequently needs to get update to maintain the node state. If any node state information wrong which impacts on overall network route

(continued)

(continued)

Author	Title	Research methodology	Research gap analysis
Amjad Mahamood (Feb 2017) [10]	ELDC: An Artificial Neural Network-based Energy-Efficient and Robust Routing Scheme for Pollution Monitoring in WSNs	Amjad et al. proposed a group based protocol based on the dynamic cluster process. The evaluation of dynamic cluster formation based on the network condition improves the node selection and route route selection. To minimize the energy consumption of the network this mechanism takes the consideration of EEUC (energy-efficient unequal clustering). This process assigns border group ahead to distribute the group data to the other group users	The dynamic cluster formation consumes more energy for organizing group-based communication
Sudeep Tanvar, Sudanshu Thyagi (2018) [11]	LA-MHR: Learning Automata Based Multi-level Heterogeneous Routing for Opportunistic Shared Spectrum Access to Enhance Lifetime of WS	The LA-based multihop heterogeneous routing improves the sensing node stability by validating the sensing field. The cluster head selection process evaluated based on the spectrum data and based on the SLA. The BS allocates a spectrum to the selected CH. The spectrum allocation rate estimated based on the distance of Base Station to the CH distance rate	The major limitation here is a multi-level process which takes more travelling cost
Xi Tao and Wei Song, Senior Member, IEEE (2018) [12]	Location-Dependent Task Allocation for Mobile Crowd sensing with Clustering Effect	This paper discovers resource allocation problem from two different features. The first process focuses on data distribution and designs a genetic algorithm (GA) to maximize data distribution quality. Then, the next process considers the profit of nodes into account and proposes a detective algorithm (DA) to improve the profit	Major problem is resource allocation due to the lack of node cooperative communication problem

(continued)

(continued)

Author	Title	Research methodology	Research gap analysis
Tung-Wei Kuo, Kate Ching-Ju Lin, and Ming-Jer Tsai [13]	On the Construction of Data Aggregation Tree with Minimum Energy Cost in Wireless Sensor Networks: NP-Completeness and Approximation Algorithms	This paper presents the problem of data aggregation problem. This problem was overcome with relay node NP-complete. Using NP-complete formation the approximate route formation will be estimated and process the data aggregation	This paper resolves the data aggregation construction process by using NP-Completeness. But this process needs more computation resources and requires more energy
Mohammed Mohsen Mohammed Nasr, Abdeldime Mohamed Salih and Lian-Feng Shen (2016) [14]	Analytical Exploration of Energy Savings for Parked Vehicles to Enhance VANET Connectivity	This paper focuses on energy-saving process of VANET model by considering the relay node energy state and by discovering the optimal energy harvested relay nodes. In this process, first, the relay nodes and forwarder nodes are elected based on the energy and distance rate values and identifies the current load of relay nodes it distributes the resources and workload	If more number of vehicle increase more number of relay and forward nodes required
Andrei Horvat MaLevente, Fuksz Petrică C. Pop and Daniela Dănciulescu (2015) [15]	A Novel Hybrid Algorithm for Solving the Clustered Vehicle Routing Problem	This paper designs a hybrid route optimization method based on the genetic algorithm to resolve the clustering and routing process problem. The protocol determines an NP-hard combinatorial optimization problem that generalizes the classical vehicle routing problem (VRP) and generalized vehicle routing problem (GVRP)	The hybrid novel approach overcomes the routing problem but failure to achieve network lifetime

3 Network Modelling

We considered a set of sensor nodes with different states such as handoff state and forwarding state. All the sensor nodes are distributed through the network and can initiate communicate any arbitrary directions with the minimum energy rate of λ_{E_i} and handoff rate λ_{o_i} . The initial sensing, transmission and receiving rate defined as $\{\lambda_{E_s}, \lambda_{E_{tx}}, \lambda_{E_{rx}}\}$. Initial communication service rate defined as μ_i . The network is

divided into a set of clusters $\{C_i\}$, each cluster organizes the set of sensor nodes with initial communication rates. The following equation organizes set of clusters for set of nodes and with their corresponding communication states.

$$C_i = \frac{|n \log_{10} N|}{\mu} \quad (1)$$

where C_i represents a set of clusters in a network, n is set of nodes and N is a total network area

Each node state in each cluster varied with respective of energy rates and communication rates, we adopt Hidden Markov model (HMM) to analyse each sensor state by estimating each node energy rate λ_{E_n} and communication propagation rate μ_{n_i} . In this HMM mode, each sensor node states is represented as $s_i = \{n_s, n_{tx}, n_{cs}, n_{idle}\}$, the node state will transfer from one state to another statue in cluster state and following model represents the transition model to determine and analyse the sensor node state level.

The following definitions are determined to analyse the sensor node state.

Definition 1: Busy state To determine the node busy state $P(n_b)$, we estimate the node busy state probability by considering average difference rate of node communication arrival rate λ_o and handoff rate λ_H .

Definition 2: Transmission state The transmission state of sensor node $p(n_{tx})$ derived based on the estimation of current node communication range rate μ and average distance rate μ_d .

Definition 3: Handoff state The sensor node handoff state probability determines the probability node current energy rate λ_{E_n} and communication distance rate μ_d of current sensor node.

The symmetric equation for node state $\{s_i\}$ is derived as

$$\sum_{i=0}^s P(s_i) = \frac{\lambda_o + \lambda_H}{i\mu} P(i-1), \quad 0 \leq i \leq S. \quad (2)$$

The average rate of all nodes states must be equal to one:

$$\sum_{i=0}^S P(s_i) = 1. \quad (3)$$

The communication blocking probability B_O when all S sensor nodes are busy, which it can be derived as

$$B_O = P(S) = \frac{\frac{(\lambda_o + \lambda_H)^S}{S! \mu^S}}{\sum_{i=0}^S \frac{(\lambda_o + \lambda_H)^i}{i! \mu^i}} \quad (4)$$

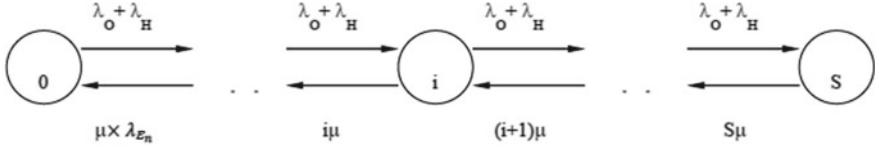


Fig. 2 Node state transition model

The node balance state equations derived as

$$\begin{cases} i\mu P(i) = (\lambda_O + \lambda_H)P(i-1), & 0 \leq i \leq S_C \\ i\mu P(i) = \lambda_H P(i-1), & S_C \leq i \leq S \end{cases}. \quad (5)$$

The average rate of the overall state is

$$\sum_{i=0}^S P(i) = 1. \quad (6)$$

The blocking probability $P(B_o)$ for organizing communication is derived as (when a set of sensor nodes S_c are busy state):

$$P(B_o) = \sum_{i=S_c}^S P(i). \quad (7)$$

The blocking probability $P(B_H)$ for a handoff communication is when a set of S sensor nodes busy in a cluster S_c (Fig. 2)

$$P(B_H) = P(S) = \frac{(\lambda_O + \lambda_H)^{S_c} \lambda_H^{S-S_c}}{S! \mu^S} P(0). \quad (8)$$

3.1 Adaptive Energy Rate Allocation

Wireless nodes have low data error rates. Sensed data is compressed and available at required data rates. By maintaining the fixed data rate, error will be low and leads to require that physical layer sensor node information be made available at the MAC layer. The proposed method is to solve data rate fluctuation, which makes the use of state information of sensor nodes. This makes to determine transmission data rate. By using sensor node estimation power indication P_r , can be calculated as

$$P_r = \frac{P_t \times G_t \times G_r \times H_t \times H_r \times \lambda^2}{(4 \times \pi \times d)^2 \times L} \quad (9)$$

Signal power at transmitting node is represented with P_t , and signal power at receiving node is represented with P_r . The free space propagation model is represented with Eq. (2). Transmitter gain is represented with G_t and receiver gain is represented with G_r , H_t and H_r are height of the transmitter and receiver, λ is wavelength, d is distance between the transmitter and receiver and L is system loss. The transmission data rate is mapped by the received signal strength. This data rate matching is done by threshold-based technique. Receiver sends data to transmitter in a determined bit rate. By receiving the data rate of transmitter, the receiver adjusts the data rate accordingly at the physical layer. Other neighbour nodes that hear the packet will update the information in their network allocation vector (NAV) and hold their transmission until current transmission gets completed.

To minimize the energy consumption, an energy allocation scheme is designed. Resource allocation, joint power control, scheduling schemes over the time window T can be expressed as

$$\text{Minimize} \sum_{m=1}^M \sum_{t=1}^T p_t^m \quad (10)$$

$$\text{subject to} \sum_{m=1}^M p_t^m \leq P_{max} \forall t \quad (11)$$

$$\sum_{t=1}^T b_m^t \log_2 \left(1 + \frac{h_t^m p_t^m}{N_o b_t^m} \right) = W^m \forall m \quad (12)$$

$$\sum_{m=1}^M b_t^m \leq B \forall t \quad (13)$$

The objective function of problem (P1) expresses the total power consumption assigned to all users across all time intervals. Constraints (9) guarantee that the power allocated to all users at each time interval is below the ceiling value P_{max} . Constraints (12) guarantee that the information message will be delivered to each user within the predefined time horizon of T seconds and (13) limits the power assigned per user at each time slot to the system power. Inequalities (14) define the continues variables of the problem. Note that in problem (P1) the non-linearities are found in the constraints of the problem. It is easy to show that this problem can be transformed into a nominal convex non-linear optimization problem by linearizing the constraints as follows:

$$\text{Minimize} \sum_{m=1}^M \sum_{t=1}^T \frac{N_o b_t^m}{h_t^m} \left(2^{\frac{r_t^m}{b_t^m}} - 1 \right) \quad (13)$$

$$\text{subject to} \sum_{m=1}^M r_t^m \leq R_t \forall t \quad (14)$$

$$\sum_{t=1}^T r_t^m = W^m \forall m \quad (15)$$

$$r_t^m \geq 0 \quad (16)$$

$$b_t^m \geq 0 \forall t, m \quad (17)$$

This gives the optimum energy rate allocation in order to achieve minimum power consumption within the message delivery delays.

For a given aggregate data requirement, $\sum_{m=1}^M W^m = W$, with delay flexibility T , the optimal rate allocation to minimize the downlink power consumption can be expressed as follows:

This gives the optimum energy rate allocation in order to achieve minimum power consumption within the message delivery delays. For a given aggregate data requirement, $\sum_{m=1}^M W^m = W$, with delay flexibility T , the optimal rate allocation to minimize the downlink power consumption can be expressed as follows:

$$\mathbf{f}(r_t) = \text{Minimize} \sum_{t=1}^T \frac{N_o b_t^m}{h_t^m} \left(2^{\frac{r_t^m}{h_t^m}} \right) \quad (18)$$

$$\sum_{t=1}^T r_t^m = W \quad (19)$$

h_t is the average sensor node gain of the moving terminals and r_t is the total rate allocated at time t to satisfy all requests. Equation (19) is monotonically increasing and convex function in r . Problem (P3) is an optimization problem over the simplex (20). Using the Karush–Kuhn–Tucker (KKT) optimality conditions, we show that this system of equations can be solved analytically for r_t . For a local minimum r_t in the system of Eqs. (19) and (20), there exists a scalar λ_- such that

$$r_t^* \in \arg \min \{f(r_t) - \lambda^* \left(\sum_{t=1}^T r_t - W \right) \} \quad (20)$$

The first-order necessary condition is $\frac{\partial f(r_t^*)}{\partial r_t} = \lambda^*$ while $\sum_{t=1}^T r_t^m = W$ Eq. (16) and λ^* is unconstraint. For these conditions, the following holds:

$$\sum_{t=1}^T B \log_2 \frac{\lambda^* h_t}{N_o \ln 2} = W \quad (21)$$

Solving for λ^* and substituting back in $\frac{\partial f(r_t^*)}{\partial r_t} = \lambda^*$, r_t^* can be derived as

$$r_t^* = \frac{W}{T} + B \log_2 \frac{h_t}{\sqrt{h_1 h_2 \dots h_T}} \quad (22)$$

4 Adaptive Genetic Corelation Node Optimization

In this research, we categorize a GA to optimize the cluster head assignment for Wireless Sensor Network by categorizing into two different steps. First, the initial population is considered as a set of cluster members. Second, a chromosome represents a set of cluster members C_i assigned to a base station B . Each chromosome represent as a $k \times k$ matrix, each row represents a set of cluster members assigned to a base station B . The chromosome of the (i, j) th element is set to 1 if he particular cluster members $\{C_i\}$ allocates to the particular base station B_i , if the cluster members are unused it set as 0. The number of 1's in each row is M and number of 0's is $M - N$. We setup the used and unused cluster members to the base station to maximize the energy and optimize the energy rate by allowing the cluster member borrowing optimization .

The detailed process is described as follows. The evaluation function $F(c, g)$, for low energy rate a chromosome c at cluster member g is

$$F(c, g) = \begin{cases} \emptyset(c, g), & \text{if } \Phi(c, g) \geq (1 - \epsilon) \cdot \Phi_{cn} \\ 0 & \end{cases} \quad (23)$$

$\emptyset(c, g)$ is the aggregate energy of low energy handlers of chromosome c at generation g as

$$\emptyset(c, g) = \sum_i \sum_{j \in C_i(c, g)} \sum_{n \in \mathcal{L}_i(c, g)} R_{i,n}^j \quad (24)$$

$C_i(c, g)$ and $\mathcal{L}_i(c, g)$ are set of chosen cluster members at base station B_i , a set of low energy operators under base station of chromosome c at generation g . $\Phi(c, g)$ is the corresponding energy for all users computed as

$$\Phi(c, g) = \sum_i \sum_{j \in C_i(c, g)} \sum_{n \in v_i} R_{i,n}^j \quad (25)$$

v_i denotes the set of all users at base station B_i , Φ_{cn} is the aggregate energy of all users using the cluster member sets determined by the adaptive rate allocation scheme in Sect. 4

Initial Population Generation

In this section, we consider a set of cluster member allocation which were determined by the conventional scheme Φ_{cn} , choose a random base station to generate a chromosome for used and unused cluster members. The used cluster members are randomly selected and replaced with unused cluster members. A chromosome is generated for each cluster member at each chosen cluster head, some used cluster members are randomly elected and substituted with unexploited cluster members. If the aggregate energy rate of the low energy sensor nodes is less than the conventional scheme function $F(c, g)$, then this chromosome is discard, and this process is repeated until a optimal chromosome is found. When such a chromosome is discovered, then another base station is randomly elected, and the entire process is reorganized until to form the initial population.

Initial population Formation

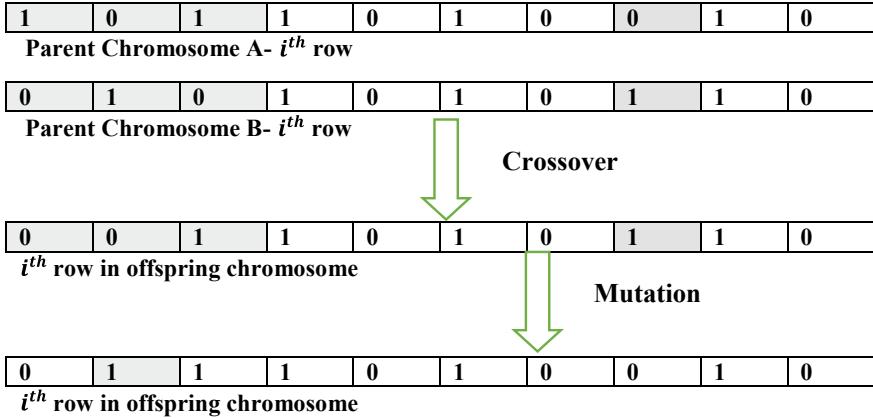
Step 1: Consider a cluster C_i with k number of calls, the cluster member frequency for the $k - 1$ calls is assigned by $f_{i*k} = (k - 1) \times \alpha + 1$, where α is the minimum frequency rate for maximum demand i^{th} cluster, and is assigned by $\alpha = \left[\frac{\beta}{m_i^*} \right]$ where β is a total number of lower bound o required frequencies in the network and $\alpha > c_i^*$

Step 2: Let discovery the next largest number of calls for the cluster C_{i-1}

- (a) Estimate a number of available frequencies in the subgroup whose size is α
- (b) Randomly choose a frequency from the frequency block which it was represented in step (a)
- (c) Assign a frequency to the randomly chosen subgroup.

Assign a frequency to the chosen subgroup for the next call, with a regular time interval with previous assigned frequencies. The assignment should satisfy the co-cluster member constraint or adjacent cluster member constraint.

Step 3: Repeat the assigning process to the remaining subgroups. Let consider the base station chromosomes information, which was estimated in chromosome generation section, based on the generated chromosome data, if the two chromosome elements have the same value, this value is assigned to the corresponding element position of the offspring chromosome. The outstanding elements of the offspring are occupied with the randomly chosen values based on this condition If $if(P_r) > P_X$ where P_r is randomly generated probability and P_X is a crossover probability, then generate a random crossover point and assign energy rate.



The above figure represents the crossover and mutation process, and based on the above figure, the total number of cluster members assigned to the base station $B = 10$ and total number of cluster members used are $Ch = 5$. Based on the figure, both parent chromosome (A and B) elements are same, where the remaining elements in offspring chromosome are filled with randomly generated elements of both parent chromosomes, where the elements 2 and 3 of parent chromosome A are copied to the offspring, as are elements 1 and 8 of parent chromosome B.

Once the crossover process evaluated the results, the mutation process initiate with the mutation probability $P(mu)$ at each row of the offspring chromosome. Randomly selected elements 1's are replaced with randomly selected elements 0's. A randomly chosen row of the corresponding offspring is replaced with the randomly generated row containing M 1's. This indicates that the cluster member set of a randomly chosen basestation is fully re-generated.

5 Genetic Corelation Node Optimization Routing

In GCNO routing algorithm, a node broadcast a route RREQ packet to discover the next corresponding node based on node intimacy and energy value features

$$R_{ij}(req) = \tau_{ij}^\alpha n_{ij}^\beta$$

Based on the below network graph, each node contains with their own node intimacy rate trails, the GCNO routing protocol broadcast GCNO RREQ packet by considering initial intimacy rate trails rate to discover shortest path-based node corelation weightage value. The initial route discovery starts from node A and it reaches destination node G (Fig. 3).

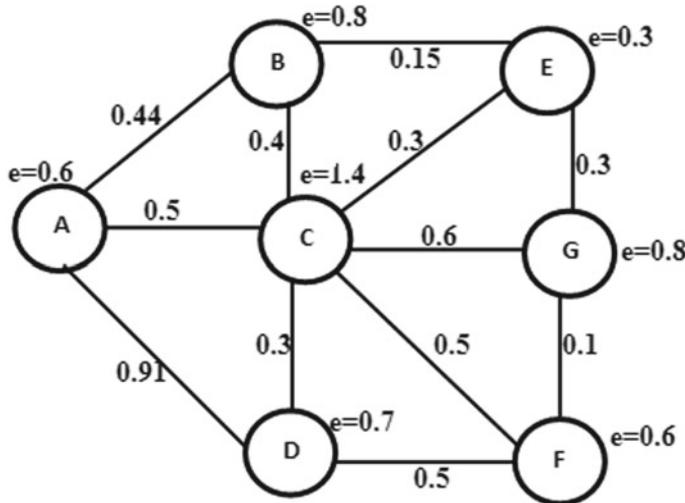


Fig. 3 Network graph

The graph can be represented in the form of matrix

$$\begin{bmatrix} & A & B & C & D & E & F & G \\ A & 0 & 0.44 & 0.5 & 0.41 & \infty & \infty & \infty \\ B & \infty & 0 & 0.3 & \infty & 0.15 & \infty & \infty \\ C & \infty & 0.4 & 0 & 0.3 & 0.3 & 0.5 & 0.6 \\ D & \infty & \infty & 0.3 & 0 & \infty & 0.6 & \infty \\ E & \infty & 0.15 & 0.3 & \infty & 0 & \infty & 0.3 \\ F & \infty & \infty & 0.5 & 0.5 & \infty & 0 & 0.1 \\ G & \infty & \infty & 0.6 & \infty & 0.3 & 0.1 & 0 \end{bmatrix}$$

Theorem 1 Considered $\{\alpha, \beta, \rho\} = \{10, 2, 1\}$.

A:

$$\begin{aligned} B(s, e) \quad \varphi_{ab} &= \frac{B_e}{B_e + C_e + D_e} \quad P_{ab} = s^\alpha(\varphi_{ab}) \\ : \quad C(s, e) \quad \varphi_{ac} &= \frac{C_e}{B_e + C_e + D_e} \quad P_{ac} = s^\alpha(\varphi_{ac}) \\ : \quad D(s, e) \quad \varphi_{ad} &= \frac{D_e}{B_e + C_e + D_e} \quad P_{ad} = s^\alpha(\varphi_{ad}) \end{aligned}$$

B:

$$C(s, e) \quad \varphi_{ac} = \frac{C_e}{C_e + E_e} \quad P_{bc} = s^\alpha(\varphi_{bc})$$

$$: E(s, e) \quad \varphi_{ac} = \frac{E_e}{C_e + E_e} \quad P_{be} = s^\alpha(\varphi_{be})$$

C:

$$B(s, e) \quad \varphi_{cb} = \frac{B_e}{B_e + E_e + G_e + F_e + D_e} \quad P_{cb} = s^\alpha(\varphi_{cb})$$

$$: E(s, e) \quad \varphi_{ce} = \frac{E_e}{B_e + E_e + G_e + F_e + D_e} \quad P_{ce} = s^\alpha(\varphi_{ce})$$

$$: G(s, e) \quad \varphi_{cg} = \frac{G_e}{B_e + E_e + G_e + F_e + D_e} \quad P_{cg} = s^\alpha(\varphi_{cg})$$

$$: F(s, e) \quad \varphi_{cf} = \frac{F_e}{B_e + E_e + G_e + F_e + D_e} \quad P_{cf} = s^\alpha(\varphi_{cf})$$

$$: D(s, e) \quad \varphi_{cd} = \frac{B_e}{B_e + E_e + G_e + F_e + D_e} \quad P_{cd} = s^\alpha(\varphi_{cd})$$

D:

$$C(s, e) \quad \varphi_{dc} = \frac{C_e}{C_e + F_e} \quad P_{dc} = s^\alpha(\varphi_{dc})$$

$$: F(s, e) \quad \varphi_{df} = \frac{F_e}{C_e + F_e} \quad P_{df} = s^\alpha(\varphi_{df})$$

E:

$$B(s, e) \quad \varphi_{eb} = \frac{B_e}{B_e + C_e + G_e} \quad P_{eb} = s^\alpha(\varphi_{eb})$$

$$: C(s, e) \quad \varphi_{ec} = \frac{C_e}{B_e + C_e + G_e} \quad P_{ec} = s^\alpha(\varphi_{ec})$$

$$: G(s, e) \quad \varphi_{eg} = \frac{G_e}{B_e + C_e + G_e} \quad P_{eg} = s^\alpha(\varphi_{eg})$$

F:

$$D(s, e) \quad \varphi_{fd} = \frac{D_e}{D_e + C_e + G_e} \quad P_{fd} = s^\alpha(\varphi_{fd})$$

$$\begin{aligned} & : C(s, e) \quad \varphi_{fc} = \frac{C_e}{D_e + C_e + G_e} \quad P_{fc} = s^\alpha(\varphi_{fc}) \\ & : G(s, e) \quad \varphi_{fg} = \frac{G_e}{D_e + C_e + G_e} \quad P_{fg} = s^\alpha(\varphi_{fg}) \end{aligned}$$

Case 1: let assume $\{\alpha, \beta, \rho\} = \{5, 3, 5\}$.

Table 1 represents the visited and unvisited nodes and functional values, each packet bounded with bit vector [0, 0, 0, 0, 0, 0, 0] to represent visited and unvisited information. According to Table 2, node A elects node C over node B, due to the higher intimacy and energy rate function value of AC. Then the bit vector of C value set as 1 will be considered as higher energy rate node, rather than other nodes. From node C, to next hop node B, based on the higher intimacy and energy value function, and set a bit vector as 1 and CB is considered because it has higher energy rate compared to CD. From node B, the to next node E, based on the route discovery rate function, and set a bit vector as 1 and BE is considered because it has higher intimacy and energy value compared to other nodes. From node E to the route discovery process discover the destination G and finds the destination G and elects the optimal path EG and set E and G bit vector value as 1. Finally, the source node A reaches to the destination node through optimal path of $A \rightarrow C \rightarrow B \rightarrow E \rightarrow G$.

5.1 RREP Fault Diagnosis Routing Vector at Destination

Based on the Case 1, three consecutive bits are used to discover the optimal routing path, based on the previous case; the node A begins the route discovery process with a set of bit vectors in route vector. Based on the final routing vector set [011010100000000000000000] bits, the first three bits 001 be examined as node C, next three bits 011 considered as node B, and next three bits 010 considered as node E and next three bits 100 as destination node G. Finally, the discovered optimal path is A-C-B-E-G.

Case 2: let assume $\{\alpha, \beta, \rho\} = \{5, 2, 3\}$.

In order to evaluate the corelation process at destination, an GCNO RREP routing vector is considered. In this expertise, the route vector initiates the route bit vector at destination and process the higher intimacy and energy rate function, to elect the optimal nodes. According to Table 2, the destination node G chooses C node over E, C because of its higher rate and establish a path of $C \leftarrow G$ over GE and GB. Now, the node C elects next higher energy node as B over D and establish a path of $B \leftarrow C \leftarrow G$. Finally from node B, identifies the target source node, and establish the optimal and corelation RREP as $A \leftarrow B \leftarrow C \leftarrow G$. Based on this trusted path, the bit vector gets changed.

Table 1 GCNO RREO route

Table 2 GCNO RREP

6 Experimental Study

In this section, we analyse the performance of the proposed GCNO scheme. We simulated WSN with a set of sensor nodes and each mobile is represented as a sensor device, which captures a data and transfer data towards destination. We compared the performance of Genetic Corelation Node Optimization (GCNO) on these parameters packet delivery ratio (PDR), Average throughput, Average delay, energy consumption and network overhead. We compare the performance of GCNO with particle swarm optimization based energy-efficient cluster head Selection algorithm [16]. The proposed system is simulated with the network simulator-2 (NS-2) [17] with the simulation parameters of Table 1.

No. of nodes	300, 400, 500 and 600
Area size	1000×1000
Mac	802.11
Routing protocol	GCNO
Transmission range	250 m
Simulation time	20 s
Traffic source	CBR
Packet size	512
Receiving power	0.395
Sending power	0.660
Idle power	0.035
Initial energy	10.0–50 J
Data rate	2 Mbps

6.1 Simulation Results

In this experimental model, we simulate the WSN model with the variation of number of nodes and energy levels. In this simulation we consider the network area size as $1000 \text{ m} \times 1000 \text{ m}$, for 300–600 sensor nodes, with the initial energy rate of 10 J–50 J with different number of clusters and cluster heads. Initially the nodes were dynamically placed and scattered in random locations. We compute the total number of clusters and total number of optimal cluster heads using Genetic Corelation Node Optimization scheme. We evaluated the proposed Genetic Corelation Node Optimization performance by conducting multiple simulations by varying number of clusters and cluster heads for group of nodes. First, we deploy the Genetic Corelation Node Optimization scheme to validate the performance to measure the energy

consumption rate, throughput rate, delay rate and packet delivery ratio. Initially we consider 300 sensor nodes with the energy rate between 10 and 50 J. The following results define the comparison of GCNO and PSO-ECHS.

6.1.1 Based on Number of Nodes

In this scenario, we consider different network size, we varied number of sensor nodes from 300 to 600 nodes, we have considered minimum energy rate as 10 J, we vary the number of cluster and cluster head to analyse the performance of GCNO and PSO-ECHS schemes. Based on the simulation experiments, we evaluated the performance of both schemes under different clustering environment.

Figure 4 shows the packet delivery ratio of GCNO and PSO-ECHS techniques for different number of nodes scenario. We can conclude that the packet delivery ratio of our proposed GCNO approach has 8.1% of higher than PSO-ECHS approach.

Figure 5 shows the average overhead of GCNO and PSO-ECHS techniques for different number of nodes scenario. Based on the simulation results the average overhead rate of PSO-ECHS increased with number of clusters for more number of nodes compare to proposed GCNO approach.

Figure 6 shows the energy consumption of GCNO and PSO-ECHS techniques for different number of nodes scenario. According to the results, the energy consumption of our proposed GCNO approach has lesser energy consumption compare to PSO-ECHS approach.

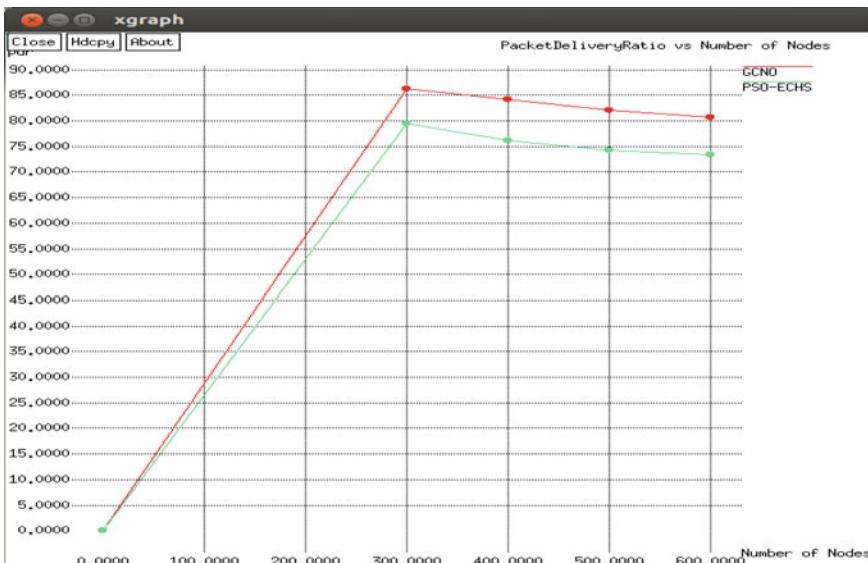


Fig. 4 Number of nodes versus packet delivery ratio



Fig. 5 Number of nodes versus overhead

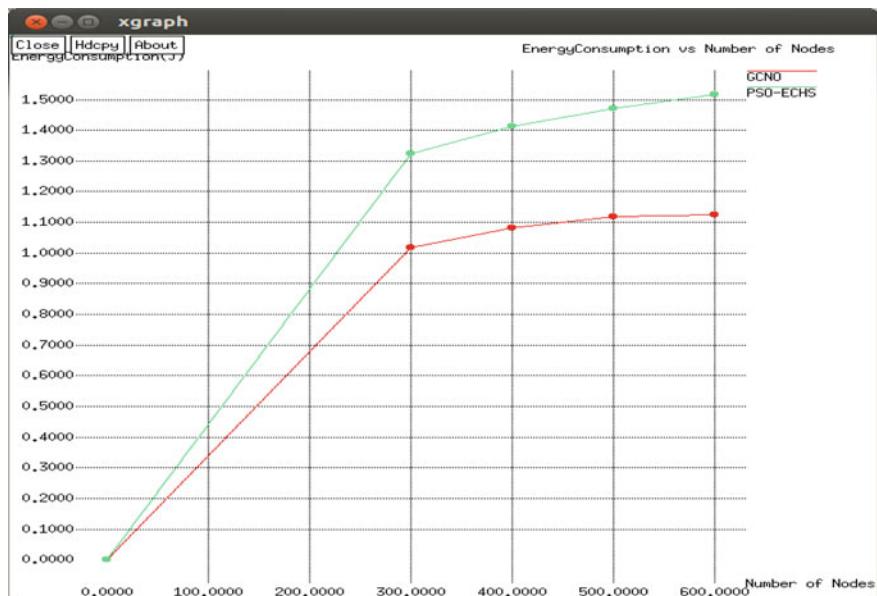


Fig. 6 Number of nodes versus energy consumption

Figure 7 shows the end-to-end delay of GCNO and PSO-ECHS techniques for different number of nodes scenario. The delay rate was increased while number of nodes increased in both schemes while comparing to GCNO the PSO-ECHS have higher delay for number of nodes.

6.1.2 Based on Energy Rates

In our second experiment, we vary the energy rate as 10, 20, 30, 40 and 50.

Figure 8 shows the packet delivery ratio of GCNO and PSO-ECHS techniques for different energy rates. We can conclude that the packet delivery ratio of our proposed GCNO approach have better packet delivery ratio compare to PSO-ECHS approach, it shows the various of 8.7% variation on both scenarios.

Figure 9 shows the average overhead of GCNO and PSO-ECHS techniques for different energy rates. Based on the results, we can observe that the overhead of our proposed GCNO approach has 7.6% of lesser than PSO-ECHS approach.

Figure 10 shows the energy consumption of GCNO and PSO-ECHS techniques for different energy rates. We can conclude that the energy consumption of our proposed GCNO approach has 10.4% of less than PSO-ECHS approach.

Figure 11 shows the end-to-end delay of GCNO and PSO-ECHS techniques for different energy rates. We can conclude that the delay in our proposed GCNO approach has 11% of less than PSO-ECHS approach.



Fig. 7 Number of nodes versus end-to-end delay



Fig. 8 Energy rate versus packet delivery ratio

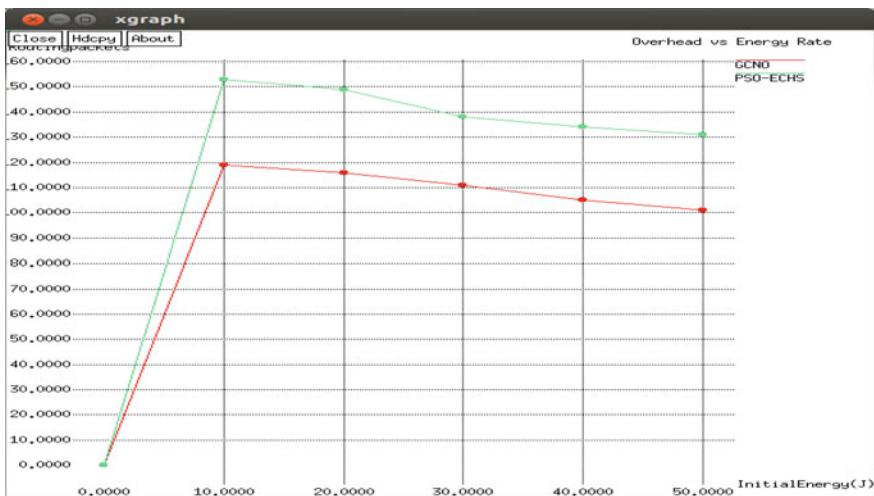


Fig. 9 Energy rate versus overhead

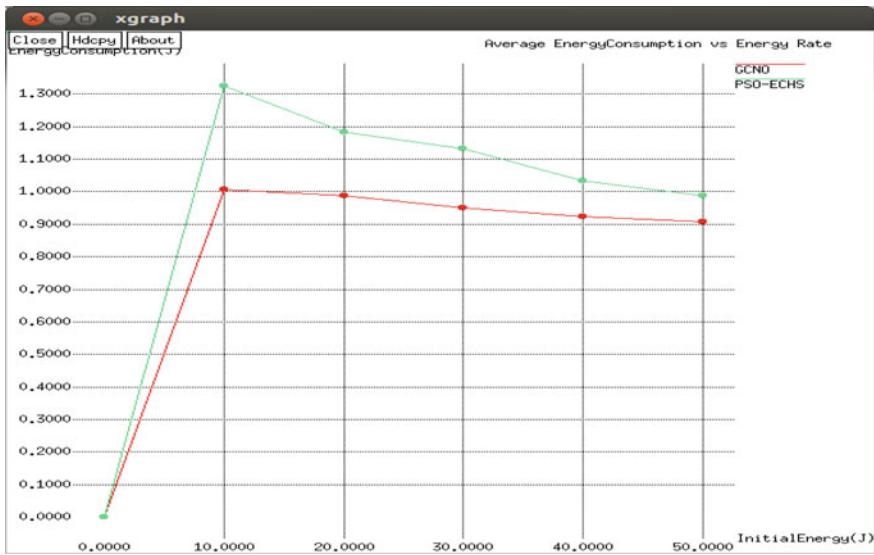


Fig. 10 Energy rate versus energy consumption

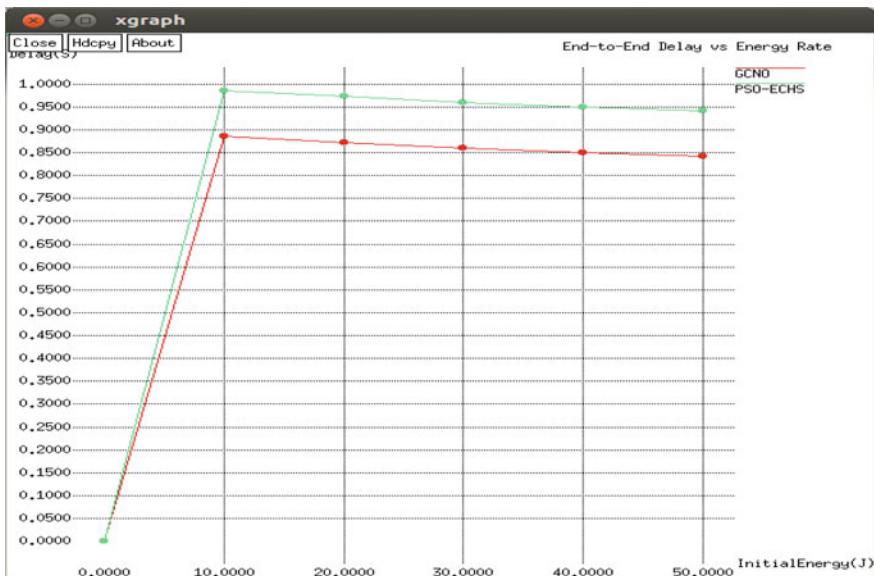


Fig. 11 Energy rate versus delay (S)

7 Conclusion

In this paper, we propose an adaptive Genetic Corelation Node Optimization Routing (GCNO) for Wireless Sensor Network (WSN). The adaptive energy rate allocation scheme minimizes the error rate and allocates the optimized clusters based on the node state. In order to identify the node fitness status, the adaptive Genetic Corelation Node Optimization scheme finds the appropriate cluster set to sort out the cluster head selection problem from a set of used and unused cluster members to maximize the data handling performance. The Genetic Algorithm approach was employed to determine the efficient energy for high energy efficient and radio access network model, and based on the simulation results, the energy is saved up to 11.91%, with increase of sensor nodes. This is achieved by designing a traditional cluster-based wireless network model by adopting traditional adaptive GA node optimization routing scheme. The proposed scheme is compared with various traditional schemes in different aspects like Energy consumption, Packet delivery ratio, Lifetime, etc.

References

1. I.F. Akyildiz et al., A survey on sensor networks. *IEEE Commun. Mag.* **40**(8), 102–114 (2002)
2. Q. Zhang et al., The design of hybrid MAC protocol for industry monitoring system based on WSN. *Procedia Eng.* **23**, 290–295 (2011)
3. A.H. Abbasi et al., Survey on clustering algorithms for wireless sensor networks. *Comput. Commun.* **30**, 2826–2841 (2010)
4. W.R. Heinzelman, A. Chandrakasan, H. Balakrishnan, Energy-efficient communication protocol for wireless microsensor networks. In *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, vol. 2 (2000), p. 10
5. S. Lindsey, C.S. Raghavendra, PEGASIS: power-efficient gathering in sensor information systems. In *Proceedings of the IEEE Aerospace Conference Proceedings*, Big Sky, MT, USA, vol. 3 (9–16 March 2002), p. 3
6. A. Manjeshwar, D.P. Agrawal, TEEN: a routing protocol for enhanced efficiency in wireless sensor networks. In *Proceedings of the 15th International Parallel and Distributed Processing Symposium*, San Francisco, CA, USA (23–27 April 2001), pp. 2009–2015
7. A. Manjeshwar, Q.-A. Zeng, D.P. Agrawal, An analytical model for information retrieval in wireless sensor networks using enhanced APTEEN protocol. *IEEE Trans. Parallel Distrib. Syst.* **13**(12), 1290–1302 (2002)
8. S. Wang, T.L.N. Nguyen, Y. Shin, Energy-efficient clustering algorithm for magnetic induction-based underwater wireless sensor networks. *IEEE Access*. <https://doi.org/10.1109/access.2018.2889910>
9. Y. Zhou, S. Taneja, C. Zhang, X. Qin, GreenDB: Energy-efficient prefetching and caching in database clusters. *IEEE Trans. Parallel Distrib. Syst.* <https://doi.org/10.1109/tpds.2018.2874014>
10. A. Mehmood, Z. Lv, J. Lloret, M. Munee Umar, ELDC: an artificial neural network based energy-efficient and robust routing scheme for pollution monitoring in WSNs. *IEEE Trans. Emerg. Top. Comput.* <https://doi.org/10.1109/tetc.2017.2671847>
11. S. Tanwar, S. Tyagi, N. Kumar, M.S. Obaidat, LA-MHR: learning automata based multilevel heterogeneous routing for opportunistic shared spectrum access to enhance lifetime of WSN. *Digit. Object Identifier*. <https://doi.org/10.1109/jstst.2018.2818618>

12. X. Tao, W. Song, Location-dependent task allocation for mobile crowdsensing with clustering effect. IEEE Internet Things J. <https://doi.org/10.1109/jiot.2018.2866973>
13. T-W. Kuo, M-J. Tsai, On the construction of data aggregation tree with minimum energy cost in wireless sensor networks: NP-completeness and approximation algorithms. <https://doi.org/10.1109/infcom.2012.6195659>
14. M. Mohammed Nasr, A.M.S. Abdelgader, L-F. Shen, Analytical exploration of energy savings for parked vehicles to enhance VANET connectivity. IEEE Trans. Intell. Transp. Syst. (Early Access)
15. A.H. Marc, L. Fuksz, P.C. Pop, D. Dănciulessu, A novel hybrid algorithm for solving the clustered vehicle routing problem. In *Hybrid Artificial Intelligent Systems*, ed. by E. Onieva, I. Santos, E. Osaba, H. Quintián, E. Corchado. HAIS 2015. Lecture Notes in Computer Science, vol. 9121 (Springer)
16. P.C. Srinivasa Rao, P.K. Jana, H. Banka, A particle swarm optimization based energy efficient cluster head selection algorithm for wireless sensor networks (Springer Science+Business Media New York, 2016)
17. L.F. Akyildiz, T. Melodia, K.R. Chowdhury, A survey on wireless multimedia sensor networks. Comput. Netw. (Elsevier) **51**(4), 921–960 (2007)

Nandoori Srikanth is one of the part time Ph.D. scholars in Koneru Lakshmi Educational Foundation, from the Department of Electronics and Communication Engineering. He published many research papers in various reputed journals and he is working as an Assistant professor in NRI Institute of Technology. His research area is Wireless Sensor Networks. His research interests are wireless communications and signal processing.

Muktyala Siva Ganga Prasad is one of the Professors in Koneru Lakshmi Educational Foundation, from the Department of Electronics and Communication Engineering. He published many research papers in various reputed international journals and he guided many more researchers in the fields of wireless communication, antennas and wireless sensor networks. His research interests are wireless communications, antennas and signal processing.

A Novel Hybrid User Authentication Scheme Using Cognitive Ambiguous Illusion Images



Sumaiya Dabeer , Mahira Ahmad, Mohammad Sarosh Umar and Muneeb Hasan Khan

Abstract Text-based passwords are most common and easy to use but are difficult to memorize and remember. Moreover, they are prone to attacks like shoulder surfing and brute-force. On the other hand, graphical passwords are easy to remember and memorize. But they are still not commonly used as they have some issues like increased user login time, and small password space. In today's scenario where number of data breaches is increasing, more secure authentication schemes are needed to ensure the authenticity of a user. In this paper, we propose a novel hybrid user authentication scheme by integrating both text-based and graphical password schemes to make authentication system stronger and resistant to attacks. Our scheme has two steps of authentication, in which at the first step, the user has to recognize and select his appropriate image among the blurred images and in the next step, the user has to enter the tag associated with the selected image. Only after successful completion of the two steps, the user is authenticated. The images used as a part of graphical password scheme are cognitive ambiguous illusion images. The basic idea behind using these images is that they are perceived by different users differently depending on how they visualize the image. To evaluate the effectiveness of the proposed scheme, an experiment was conducted on the setup and the results obtained were promising.

Keywords Graphical password · Illusion · User authentication · Challenge response

S. Dabeer () · M. Ahmad · M. Sarosh Umar · M. Hasan Khan

Department of Computer Engineering, Zakir Husain College of Engineering and Technology, Aligarh, India

e-mail: sumaiyadabeer@zhcet.ac.in

M. Ahmad

e-mail: mahiraamu@gmail.com

M. Sarosh Umar

e-mail: saroshumar@zhcet.ac.in

M. Hasan Khan

e-mail: muneebhkhan@zhcet.ac.in

1 Introduction

In today's scenario where number of data breaches and threat to information security is increasing, much stronger and secure user authentication methods are needed to allow only authorized users to access sensitive information. Without a secure authentication system, your data, device, system, or the whole organization could be at risk. Authentication methods should be user-friendly, which means they should not put much cognitive load on the user, but meanwhile should be secure and less susceptible to security attacks.

Traditional text-based passwords are most commonly used for authentication. But it is a natural tendency of a user to choose a short and easy password and use that same password for other authentication systems too. This makes them susceptible to attacks like brute-force attacks, and dictionary attacks. But if a user chooses a difficult alphanumeric password, then it becomes hard to remember. On the other hand, graphical passwords solve the issue of remembering the passwords because humans are good at remembering and recognizing images than texts [1, 2]. Graphical passwords are a promising alternative to text-based passwords but they have their own drawbacks. Graphical passwords are susceptible to observation attacks, have low password space and slow login times [2]. A detailed discussion on Graphical Passwords is presented in [3]. To overcome the drawbacks of both the password schemes and to develop an authentication method that is less vulnerable to security attacks, the proposed approach of user authentication integrates both the text-based and graphical password schemes. It is a two-step challenge response-type authentication. The first authentication step incorporates the graphical password scheme while the second step of authentication is text-based. Only after the successful completion of both the steps, a user is authenticated. The images used in the first step are cognitive ambiguous illusion images, which are perceived by different users differently. Ambiguous illusions are pictures or objects that elicit a perceptual switch between the alternative interpretations. A popular example is of Rubin Vase with two visual representations.

The remaining paper is structured as follows: Sect. 2 sums up the literature review of this field. Section 3 describes the proposed method along with the example. Method analysis and User study are presented in Sects. 4 and 5 respectively. Finally, Sect. 6 concludes the work followed by Future work in Sect. 7.

2 Background Study and Related Work

A lot of notable work is already done in graphical and text based password schemes. Some of them are listed below.

In Use Your Illusion by Pering et al. [4], a user is allowed to select his own set of images as passwords. Then the images are distorted and in the training phase, the original images and the distorted images are shown to the user to memorize. During

login, user has to select his distorted images among the other distorted images. But this scheme suffers from low password space.

The authors K. Divyapriya and Dr. P. Prabhu in [5] proposed an authentication scheme for touch screen devices to resist shoulder surfing attack, taking into account that an attacker observes the screen at a distance. The virtual keypad of the Illusion-Pin is a combination of two keypads with different digit orderings. Thus, a person close to the screen will see one keypad while the shoulder surfer, at a distance, will be able to see only the other keypad.

Welch et al. in [6] proposed an improvement over the traditional passfaces scheme. In this scheme, some alphanumeric characters are associated with each image. The user does not have to click the appropriate images but enters the characters associated with them as password. This scheme, however, suffers from shoulder surfing attack and puts more strain on the user.

Hui et al. [7] proposed a conceptual framework for high-end graphical password in which after the user enters the username, the pass images are loaded and blurred immediately. The user can perform rotation and resizing functions on the images to match them against their preset angle, size, and sequence in the database.

Umar et al. in [8] proposed a graphical user authentication scheme based on the time interval that employs graphical coordinates along with a novel introduction of time interval between successive clicks. In this scheme, the user needs to recall the coordinates and the time interval between successive clicks.

Istyaq et al. in [9], Umar et al. [10], Usmani et al. [11], Saeed et al. [12], and Agrawal et al. [13] have done some recent and notable work in the field of graphical password scheme of authentication.

Zheng et al. in [14] proposed a hybrid authentication scheme based on shape and text. In this scheme, the original password uses shapes and strokes on the grid as the shape of stroke can be easier to remember than text and in the login phase, the user uses the keyboard to input the password.

Fatima et al. in [15] proposed a novel challenge response type of user authentication, which aims at providing higher level of security than conventional text-based passwords as the password change in each session of authentication, even the actual password remains the same. More notable work on text-based scheme of user authentication is presented in [16, 17].

Yu et al. in [18] proposed an evolvable graphical password authentication system: EvoPass in which a set of user selected password images are transformed into password sketches as user credentials. For user authentication, the user is required to identify and select his password sketches from a given set of challenge images. The password sketches are continually degraded to improve the password strength.

Authors in [19] proposed an alignment-based password authentication system in which for successful authentication, the user has to align the three password images in the same order as they were aligned at the time of registration.

3 Proposed Method

The proposed user authentication technique is described in this section. This scheme is a challenge response-type user authentication technique with two steps of authentication. Our proposed approach integrates both text-based and graphical password schemes to make authentication system stronger and resistant to attacks.

3.1 Register

During account setup or registration, the user will have to enter the username of his choice. Then for password setup, the system will show some optical images to the user to choose one among them which can be further used for setting up the password for that particular user. After the user chooses his desired optical image, the system will ask the user to provide his perception of that optical illusion image. Based on the user's response, the system will show some images to the user and the user will select any image of his choice, among the displayed images, as an answer to that original optical image. Once the image is selected by the user, it is blurred or distorted by some amount. This blurred image is the password image for that particular user. Blurring the image makes the scheme more resistant to shoulder surfing and observation attacks.

Finally, the original optical image and the blurred image, which the user chooses as an answer to the optical image, are shown side by side so that both the images get registered in the user's mind. This was all about setting up the graphical password.

To deal with the problem of shoulder surfing, the user needs to choose one of the methods to calculate effective position of the image to be clicked with respect to the password image at login time, so that clicked image changes at each login but the user's password image remains same. These operations may be like transpose position or diagonally left, diagonally right or any combination of displacement. Now to overcome the issues related to the graphical passwords like low password space, we are also incorporating text-based password scheme into the proposed scheme. After the blurred image is set as password image for the user, the user is asked to set a tag of not less than eight characters for the chosen password image. The tag can contain alphabets or numbers, special characters or combination of all three. It will be easier for the user to remember and recall the tag on seeing the corresponding password image.

3.2 Login

Login phase consists of two steps of user authentication. First, the user enters the username. Only after successful validation and verification of the username, the user

is directed to the password verification steps. In the first authentication step, the original optical illusion image is displayed on the screen along with nine blurred or distorted images. Among the nine blurred images, there will be only one correct password image. The original optical illusion image is shown as a hint to the user to recall his password image. Moreover, the user can recognize his blurred password image among the other decoy images by using color and shape cues. But the user need not to click his blurred image directly as it may lead to shoulder surfing attack. User calculates the effective location to be clicked by recalling the method he selected during registration.

Result: User will be registered if details are ok

Ask user to provide Username;

if *Username is available then*

 Display optical image OI and record perception;

 Display images according to perception;

 Ask to choose one image I;

 Distort and Blur image DI and show alongside I;

 Ask user to choose method to calculate effective location of image to be clicked wrt password image DI;

 Ask user to enter tag for image;

 Record above details and image in database;

 Display User "registered successfully";

else

 Ask user to enter another username;

 Start again;

end

Algorithm 1: Procedure for Registration

For example, there are nine images on the login screen shown in Fig. 1, including the password image F of the user. If at the time of registration, the method selected by

Fig. 1 Example to get effective location

A	B	C
D	E	F
G	H	I

the user is Transpose, then he will select image H, which is at the transpose location of his password image F. Similarly, if the method selected was Up, then the user will have to select image C as the answer.

In the second step of authentication, the selected blurred image is displayed and the user is asked to enter the tag for that image. Seeing the image, the authentic user will be able to recall the tag associated with it. Only after the successful completion of the two steps, the user is authenticated.

3.3 Example

To get a better understanding of our proposed scheme, we are illustrating an example. For our example, we are using an optical illusion image which is shown in the Fig. 2. This picture is shown to the user at the time of registration. This optical illusion image can be perceived in two ways: some may see the face of a beautiful lady and some may see an old grumpy man. At the time of registration, suppose the user sees the face of a beautiful lady, then based upon his response, the system shows nine images of different beautiful ladies. The user selects one of the images, which is then distorted. This distorted image becomes the password image of the user. Further, the user sets the tag associated with it and selects the operation that will be used to calculate the



Fig. 2 Optically illusive image

Fig. 3 Login screen**Log In**

Username:

Don't have an account? [Click here](#) to register.

effective position of the image to be clicked, with respect to the password image, at the time of login.

In the login phase, when the user gives the url of required page and if there is no active session it will automatically redirect to login page which is depicted in Fig. 3. The user is first asked to enter the username. Only after successful validation and verification of the username, the user is redirected to the next step of authentication. In the next step, the original optical illusion image of the old grumpy man or a beautiful lady along with nine other distorted images are shown to the user.

The distorted images include the password image of the user, some images of different beautiful ladies and some images of different old grumpy men. The user is asked to tick on the effective image with respect to the password image, according to his previously selected method. This is shown in the Fig. 4. For example, if user's password image is middle one and user previously selected the diagonally left operation, then the user selects the image shown in Fig. 4. This step acts as Graphical password type of user authentication.

After ticking on the desired image, the user clicks submit button and is redirected to the next page where he is required to enter the tag associated with it. This is shown in Fig. 5.

After successful completion of both the steps, the user is authenticated.

Result: User will be authenticated if details provided are valid

Ask user to enter his username;

if *username Exists then*

 Display OI and grid of nine distorted images (one of them is DI);
 ask user to select effective location in grid;

if *inverse of user effective method(selected image)== DI then*

 Ask user to enter his tag;

if *tag is correct then*

 Authenticate user and create session;
 Display "successfull login";

else

 Authentication failed!;

end

else

 Authentication failed!;

end

else

 | Display message that user does not exist;

end

Algorithm 2: Procedure for Login

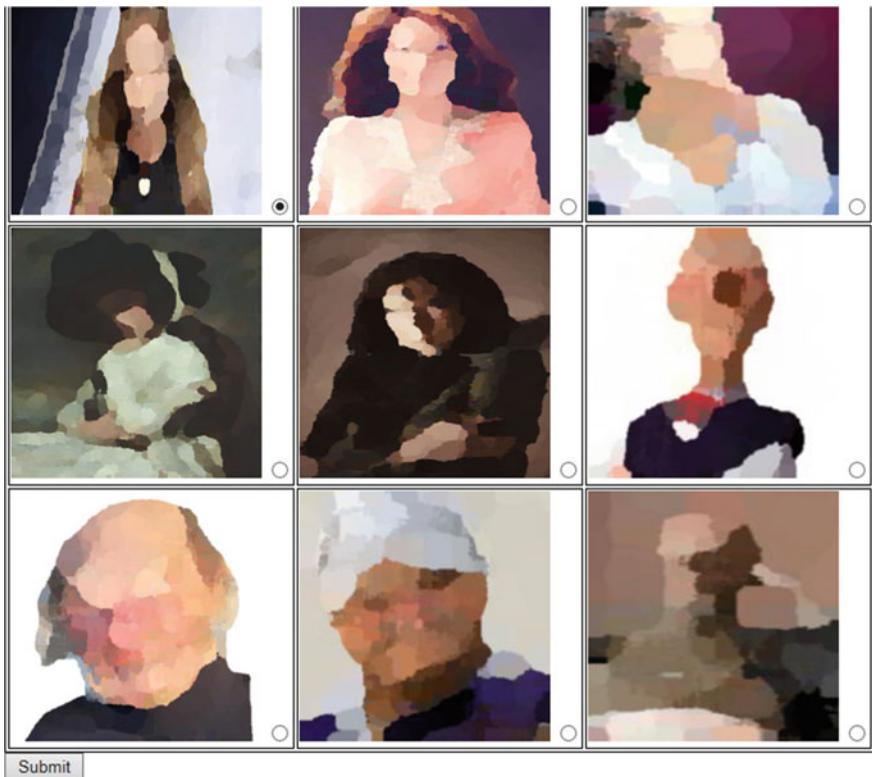


Fig. 4 First step of authentication



Fig. 5 Second step of authentication

4 Method Analysis

Both the text-based and graphical password schemes have their own benefits and drawbacks. Our proposed authentication approach combines the advantages of both the password schemes to make authentication as much resilient to attacks as possible.

4.1 Brute-Force Attack

This attack can be overcome by increasing the password space. The text-based passwords have a password space of 94^N , where N is length of the password. The pass-

word space of the graphical passwords can be enlarged by increasing the number of images. Our proposed scheme is the combination of both the text-based and graphical password schemes which eventually increases the password space. Locking the system after a number of specified attempts can also be incorporated to overcome this attack.

In the first step, the user can select correct image with probability of

$$P[\text{step1}] = \frac{1}{9} \quad (1)$$

In our case, N is atleast 8 which is length of tag in the second step. So the probability of successful attempt will be atleast

$$P[\text{step2}] = \frac{1}{94^8} = \frac{1}{6.0956893 \times 10^{15}} = 1.640503 \times 10^{-16} \quad (2)$$

Now from the multiplication rule of probability

$$P[\text{total}] = P[\text{step1}] \times P[\text{step2}] = 1.8227817 \times 10^{-17} \quad (3)$$

from the above analysis, we can deduce that this system is quite impossible to break using brute-force approach.

4.2 Shoulder Surfing Attack

The use of optical illusion images makes it difficult for an observer to understand which image is being displayed on the screen as it depends upon the user how he interprets the optical illusion images. Moreover, the relative positions of the decoy images plus the password image changes every time. Also, the blurring/distortion of the images makes it hard for the observer at a distance to clearly interpret what is being shown on the screen.

4.3 Spyware Attack

As the relative positions of the decoy images plus the password image changes every time, knowing the sequence of mouse clicks is not helpful in guessing the password image.

4.4 Intersection Attack

Intersection attack occurs when the attacker observes multiple authentication sessions of a particular user and then takes the intersection of the images to get to know about the users identity. In our scheme, we are resisting the intersection attack by always maintaining and displaying the same or identical decoy images in each authentication challenge for a particular user.

5 User Study

To evaluate the effectiveness of our proposed scheme, we conducted the usability experiment on our proposed setup. We developed a web-based authentication system which uses Django as the development environment. On login page, the user has to enter his valid username which is provided at the time of registration and hit Submit button. After successful validation and verification of the username, the user will complete the next two steps of authentication process.

User Performance: To evaluate the login time, we asked 10 different users to use our authentication system. Our usability experiment was for one week and consisted of two sessions. The two sessions included logging on the first day and after 1 week. For each successful login attempt, the user's login time is noted for both the sessions.

By analyzing the results obtained from the Table 1, we can conclude that users took longer time to login in the first session than in the second session as users were new to the system. In the second session, they took considerably less time on an average as they became familiar with the system. Moreover, this user study also

Table 1 Login time

Users	Time in sec (First day)	Time in sec (After 5 days)
User 1	20	15
User 2	28	22
User 3	15	12
User 4	30	20
User 5	22	21
User 6	21	11
User 7	19	15
User 8	16	15
User 9	26	21
User 10	32	16
Average login time	22.9	16.8

indicated that users were quick in recognizing their password image and recalling the tag associated with it.

6 Conclusion

In this paper, we proposed a novel user authentication method which combines the benefits of both the text-based and graphical password schemes and therefore is less susceptible to various security attacks. We have made use of the cognitive ambiguous illusion images. The main idea behind using such images is that these images are perceived differently by different users. We tried to illustrate how optical illusion images can be used for making more secure authentication systems. Moreover, this scheme puts less cognitive load on a user than other graphical password schemes as the user has to remember just one image and identify it. Also, remembering the tag would be easy on seeing the selected password image. At the time of user study, all the users loved working and collaborating with us as this scheme is based on hybrid method of more recognition and less recall which makes it enjoyable to user as he/she has to recall least information. In addition to this, blurred image acts as a hint for authentic user and increases the confusion for an impostor.

7 Future Work

As we proposed a relatively novel method of challenge response authentication system using a combination of both graphical and text based passwords, a lot of improvement is needed to be done for the scaling of the system. Moreover, being an authentication system it is necessary to do further enhancement on continuous basis. Some of the enhancements that can be done to make this scheme more secure and resistant to attacks are the following:

- In our method, we have only used cognitive ambiguous optical illusion images. Other cognitive illusions like distorting and paradox illusion images can be used. Physiological visual illusion images can also be used to make a stronger authentication system.
- Further experiment and analysis of the proposed system can be done with larger number of users to investigate the strength of the authentication system.
- Increasing the password space by increasing the images library.
- Using different types of filters to distort the images.
- Security of the system can be further increased by adding some extra features in text which is used as tag in our paper and by using different methods to calculate the effective position of the image to be clicked, with respect to the password image, at the time of login.

References

1. R. Dhamija, A. Perrig, (2000) Deja Vu-a user study: using images for authentication, in *USENIX Security Symposium*, vol. 9, pp. 4–4
2. E. Stobert, R. Biddle, Memory retrieval and graphical passwords, in *Proceedings of the Ninth Symposium on Usable Privacy and Security* (ACM, 2013), p. 15
3. A.V. Kayem, (2016) Graphical passwords—a discussion, in *2016 30th International Conference on Advanced Information Networking and Applications Workshops (WAINA)* (IEEE), pp. 596–600
4. E. Hayashi, R. Dhamija, N. Christin, A. Pering, Use your illusion: secure authentication usable anywhere, in *Proceedings of the 4th Symposium on Usable Privacy and Security* (ACM, 2008), pp. 35–45
5. K.D.D.P. Prabhu, Image based authentication using illusion pin for shoulder surfing attack. *Int. J. Pure Appl. Math.* **119**(7), 835–840 (2018)
6. T. Zangooei, M. Mansoori, I. Welch, A hybrid recognition and recall based approach in graphical passwords, in *Proceedings of the 24th Australian Computer-Human Interaction Conference* (ACM, 2012), pp. 665–673
7. L.T. Hui, H.K. Bashier, L.S. Hoe, G.K.O. Michael, W.K. Kwee, Conceptual framework for high-end graphical password, in *2014 2nd International Conference on Information and Communication Technology (ICoICT)* (IEEE, 2014), pp. 64–68
8. M.S. Umar, M.Q. Rafiq, J.A. Ansari, Graphical user authentication: a time interval based approach, in *2012 IEEE International Conference on Signal Processing, Computing and Control* (IEEE, 2012), pp. 1–6
9. S. Istyaq, M.S. Umar, Hybrid authentication scheme for graphical password using QR code and integrated sound signature. *Int. J. Comput. Electr. Autom. Control Inf. Eng.* 111–115 (2018)
10. M.S. Umar, M.Q. Rafiq, A graphical interface for user authentication on mobile phones, in *ACHI 2011: The Fourth International Conference on Advances in Computer-Human Interactions* (2011), pp. 69–74
11. A. Usmani, A. Maryam, M.S. Umar, M.H. Khan, New text-based user authentication scheme using CAPTCHA, in *Information and Communication Technology for Competitive Strategies* (Springer, Singapore, 2019), pp. 313–322
12. S. Saeed, M.S. Umar, PassNeighbor: a shoulder surfing resistant scheme, in *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)* (IEEE, 2016), pp. 797–802
13. S. Agrawal, A.Z. Ansari, M.S. Umar, Multimedia graphical grid based text password authentication: for advanced users, in *2016 Thirteenth International Conference on Wireless and Optical Communications Networks (WOCN)* (IEEE, 2016), pp. 1–5
14. Z. Zheng, X. Liu, L. Yin, Z. Liu, A hybrid password authentication scheme based on shape and text. *JCP* **5**(5), 765–772 (2010)
15. R. Fatima, N. Siddiqui, M.S. Umar, M.H. Khan, A novel text-based user authentication scheme using pseudo-dynamic password, in *Information and Communication Technology for Competitive Strategies* (Springer, Singapore, 2019), pp. 177–186
16. Z. Zaheer, A. Khan, M.S. Umar, M.H. Khan, One-tip secure: next-gen of text-based password, in *Information and Communication Technology for Competitive Strategies* (Springer, Singapore, 2019), pp. 235–243
17. M.H. Zaki, A. Husain, M.S. Umar, M.H. Khan, Secure pattern-key based password authentication scheme, in *2017 International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT)* (IEEE, 2017), pp. 171–174
18. X. Yu, Z. Wang, Y. Li, L. Li, W.T. Zhu, L. Song, EvoPass: evolvable graphical password against shoulder-surfing attacks. *Comput. Secur.* **70**, 179–198 (2017)
19. A. Danish, L. Sharma, H. Varshney, A.M. Khan, Alignment based graphical password authentication system, in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACOM)* (IEEE, 2016), pp. 2950–2954

Fault Classification in a Transmission Line Using Levenberg–Marquardt Algorithm Based Artificial Neural Network



Harkamaldeep Kaur and Manbir Kaur

Abstract The main objective of the power system is to supply reliable and quality electricity to all consumers. In this paper, the main focus of the author is to classify all types of faults, namely phase to ground, phase to phase, three-phase fault, and double line to ground faults that may occur at different fault locations and involve varying fault impedances in the power system using artificial neural networks (ANNs). Owing to the advantages of an artificial neural network to map nonlinearity in the data, to learn from examples and to generalize the pattern classification, ANN framework under supervised learning is implemented as a fault classifier. The proposed methodology includes extraction of features from phase voltages and currents obtained under normal and faulty conditions for different fault locations and fault impedances. The learning of feed forward ANN-based fault classifier is carried out using Levenberg–Marquardt algorithm for training the data obtained for IEEE 14 bus system.

Keywords Artificial neural network · Classifier · Faults · Levenberg–Marquardt · MATLAB

1 Introduction

Electrical faults in a transmission or distribution system occur randomly and their severity is different for different types of faults. So fault diagnoses in a power system network are essential for clearing faults that mainly occur in an electrical power transmission or distribution network. The rate of change of L-G fault is more than that of the other two phases [1]. The sharp transitions are generated under arcing

H. Kaur (✉) · M. Kaur
Thapar Institute of Engineering and Technology, Patiala, India
e-mail: harkamaldeep494@gmail.com

M. Kaur
e-mail: mkaur@thapar.edu

faults [2]. Hence, to detect and isolate the faults, a well-coordinated protection system must be provided so that the damage and disruption caused to the power system are minimized. Several protection devices are installed in the electrical power system. For example, lines are protected by the protection relays which are installed at the end and beginning of the electrical relays to detect electrical faults and switched off selectivity at the fastest possible. In the control center of the electrical power system, the task of the operator is to analyze the alarms received and this task might be difficult because of different reasons. Generally for a one-off fault, a control engineer can make a relatively swift and accurate diagnosis; however, there are times, such as stormy weather conditions, when alarm activity is very high, and this stretches and sometimes even exceeds the human ability to cope with the sheer volume of information. Multiple and interacting faults can occur at these times along with unforeseen problems in the protection mechanisms. These can severely reduce the speed of diagnosis and hence the overall efficiency. In these circumstances, there is an apparent need for an automatic, computer-based system which can be used to assist the control engineer. Such a system needs to process switching messages as they arrive to indicate the component or components involved and the type of fault which has occurred. It has to operate in real time, carrying out a diagnosis in only a few seconds and must be able to deal with all conditions including the occurrence of multiple faults or protection problems. Protection of transmission line for classification and location has been done in many research works: ANN-based fault detection and classification approach in transmission line [3], a complete protection approach for detection, classification and location in transmission line [4]. To obtain better and faster results, ANN can be implemented in relays after training and testing it [5]. ANN works like the human mind and not affected by the changes in the system parameter. The concept of fault classification and detection of faults based on the artificial neural networks using feed-forward networks and back propagation algorithm had accuracy of 86.72% [6], artificial neural network based fault Classifier (ANNFC) using discrete wavelet transform (DWT) for classification of distinctive faults on three-phase transmission line [7], the concept of fault classification in neutral non-effectively grounded distribution system using ANFIS approach and exhibited good performance at light load but accuracy was decreased at heavy load [8], fault detection classification in power system based on principal component analysis and probabilistic neural network [9], multilayer neural networks to solve the fault detection, classification and location in a transmission line system [10], fault location estimation using hybrid technique combining generalized neural network and wavelet transform [11], transmission line fault detection, and classification using Discrete Wavelet Transform [12], fault classification technique for parallel transmission line using wavelet and Clarke transformation [13].

2 Methodology

Generally, the problem of power system fault classification is solved in three stages.

Stage 1: Data is collected for different shunt types of faults on 200 km long line IEEE14 bus system using Simulink simulation. The voltages and currents of three phases are considered as fault patterns those are obtained by considering two cases

- Fault impedance
- Fault location.

Stage 2: Data is sampled to extract features from three-phase voltages and current patterns.

Stage 3: Extracted features to the classifiers for classification of faults.

2.1 System Modeling and Fault Simulation

The IEEE 14 bus system model as shown in Fig. 1 is used for collecting data using MATLAB/Simulink at various locations and on different fault impedances. The values of three-phase voltages and currents are collected as data.

Fault type: Line to ground (AG, BG, CG), Line to line (AB, BC, AC), Double line to ground (ABG, BCG, ACG), Three-phase short circuit (ABC), Three-phase short circuit with ground (ACBG).

Table 1 shows that for fault location, the model was simulated 19 times for each fault at different fault locations; therefore, total simulations for all 11 types of fault are $19 * 11 = 209$.

Similarly, for fault resistances, the model was simulated 20 times for each fault at different values of resistances; therefore total simulations for all 11 types of faults are $20 * 11 = 220$.

2.2 Data Sampling

During fault in the transmission line, a sharp change is noticed in the magnitude of voltage and current changes in the faulty phase as system transit from the normal state to a faulty state as shown in Figs. 2 and 3.

The fault is assumed to occur at the time instant 0.2 s. In total, 80,000 data points are sampled for normal and faulty conditions. For 20 cycles of input under observation, 20 points are selected randomly for three-phase voltage and current. The voltage and current values are normalized for the faulty section concerning healthy section (Tables 2, 3 and 4).

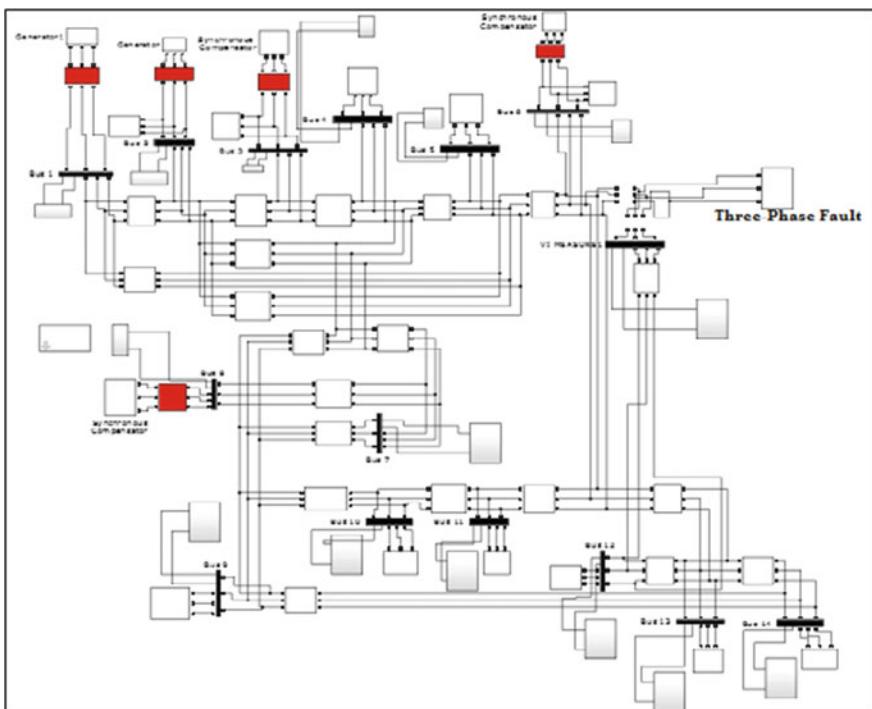


Fig. 1 Simulink model of IEEE-14 bus system

Table 1 Parameter variations

Sr. No.	Parameter	Variation values
1	Fault location (line length = 200 km) (resistance = 0.001 Ω)	10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190 (km)
2	Fault resistances (line length = 200 km)	0.001, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95 (Ω)

Fig. 2 Healthy waveform of current

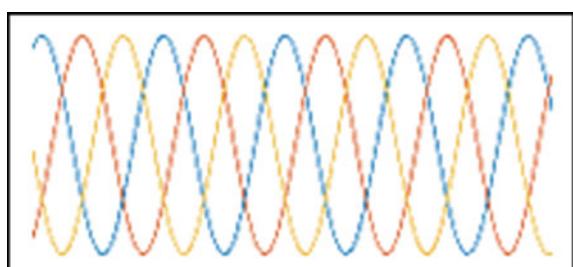


Fig. 3 Faulty waveform of current

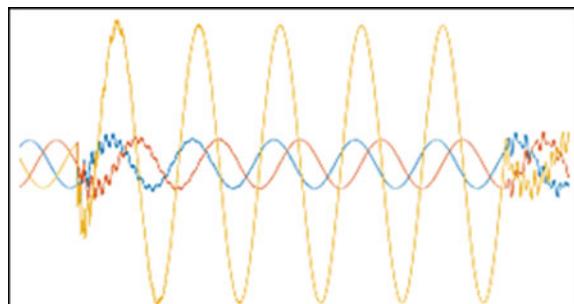


Table 2 IEEE-14 bus voltage data during normal condition

Bus no.	Bus voltage magnitude per unit
1	1.060
2	1.045
3	1.010
4	1.019
5	1.020
6	1.070
7	1.062
8	1.090
9	1.056
10	1.051
11	1.057
12	1.055
13	1.050
14	1.036

Table 3 Faulty voltage values after normalization for every cycle for ABG fault at $0.001\text{-}\Omega$ resistance

Sr. No.	V _a	V _b	V _c
1.	0.1487	0.7474	-0.8806
2.	0.1332	0.7482	-0.8952
3.	0.1578	0.7305	-0.8930
4.	0.1327	0.7545	-0.8944
5.	0.1389	0.7441	-0.8911
6.	0.1362	0.7425	-0.8774
7.	0.1489	0.7292	-0.8847
8.	0.1611	0.7402	-0.9019
9.	0.1512	0.7459	-0.8970
10.	0.1441	0.7406	-0.8847

Table 4 Faulty current (1×10^{-10}) values after normalization for every cycle for ABG fault at $0.001\text{-}\Omega$ resistance

Sr. No.	I _a	I _b	I _c
1.	-4.4	3.8	-0.11
2.	-3.8	3.4	-0.13
3.	-3.3	2.9	-0.10
4.	-3.0	2.6	-0.86
5.	-2.9	2.4	-0.88
6.	0.19	-0.093	-0.25
7.	0.003822	-0.052	0.0067
8.	0.13	-0.066	-0.048
9.	0.2	-0.024	-0.16
10.	0.1	-0.014	-0.090

Total no. of features for different fault location case = $60 * 209$

Total no. of features for different fault resistance case = $60 * 220$

2.3 Neural Network Training for Classification

There are a number of optimization algorithms, which are used to train the learning procedures in any neural network. All these algorithms have various properties in terms of performance, computational speed, and memory requirements. The most important training algorithms for neural networks are Gradient descent, Newton's Method, Conjugate gradient, Quasi-Newton, and the Levenberg–Marquardt (LM) algorithms. The LM algorithm might be best choice when there are few hundreds of parameters to train neural networks. Feed-forward neural network is selected as a classifier which is a common artificial neural network widely used to perform power system fault classification, and the Levenberg–Marquardt back propagation is chosen as a training algorithm based on mean square error. The learning process of this classifier is by updating the weight of interconnections between layers and error is calculated based on mean square error as expressed in Eq. (1).

The mean square error is calculated as

$$E(x, w) = \frac{1}{2} \sum_{p=1}^P \sum_{m=1}^M e_{p,m}^2 \quad (1)$$

where x is the input vector, w is the weight vector, m is the number of outputs and p is the number of patterns, $e_{p,m}$ is the training error at output m when applying pattern p and it is defined as in Eq. (2).

$$e_{p,m} = \text{target output} - \text{actual output} \quad (2)$$

For forward computation (input to hidden to output layers): For all layers, for all neurons in the layer. Calculate net as expressed in Eq. (3).

$$net_j = \sum_{i=1}^{ni} w_{ji} y_{ji} + \text{bias} \quad (3)$$

Calculate output as defined in Eq. (4).

$$y_j = f_j(net_j), \quad (4)$$

where f_j is the activation function of neuron j.

Bipolar Sigmoid Activation function at the input to the hidden layer is expressed in Eq. (5).

$$f(x) = \frac{1 - e^{-s_{jnet_j}}}{1 + e^{-s_{jnet_j}}} \quad (5)$$

Linear Activation function at hidden to output layer is expressed in Eq. (6).

$$f(net_j) = net_j. \quad (6)$$

Calculate slope as expressed in Eq. (7).

$$s_j = \frac{dy_i}{dnet_j} = \frac{df_i(net_j)}{dnet_j} \quad (7)$$

s_j is propagated in order, first from the inputs of the output layer to the outputs of the hidden layer, then from the outputs of the hidden layer to the inputs of the hidden layer and at the end from the inputs of the hidden layer to the input layer. This process should be repeated for other outputs.

For backward computation (output to hidden to input layers): For all outputs, i.e., for all layers for all neurons in the previous layer, for all neurons in the current layer, calculate the error.

Calculation of Jacobian matrix.

The Jacobian matrix is expressed in Eq. (8)

$$J = \begin{bmatrix} \frac{de_{i,1}}{dw_1} & \frac{de_{i,1}}{dw_2} & \frac{de_{i,1}}{dw_3} & \dots & \frac{de_{i,1}}{dw_n} \\ \vdots & \ddots & \vdots & & \vdots \\ \frac{de_{i,m}}{dw_1} & \frac{de_{i,m}}{dw_2} & \frac{de_{i,m}}{dw_3} & \dots & \frac{de_{i,m}}{dw_n} \end{bmatrix} \quad (8)$$

Size of Jacobian matrix is $p * m * n$.

p = no. of patterns, m = no. of outputs, n = no. of weights, i and j are the indices of weights from 1 to n .

The elements of the Jacobian matrix can be calculated as expressed in Eq. (9)

$$\begin{aligned}\frac{\partial e_{p,m}}{\partial w_{j,i}} &= \frac{\partial (d_{p,m} - o_{p,m})}{\partial w_{j,i}} = -\frac{\partial o_{p,m}}{\partial w_{j,i}} \\ &= -\frac{\partial o_{p,m}}{\partial y_j} \frac{\partial y_j}{\partial net_j} \frac{\partial net_j}{\partial w_{j,i}} = -F_{mj} s_j y_{j,i}\end{aligned}\quad (9)$$

where F_{mj} is the derivative of nonlinear function between output m and neuron j .

Calculation of updated weights.

The learning process of this classifier is by updating the weight of interconnections between layers as expressed in Eq. (10).

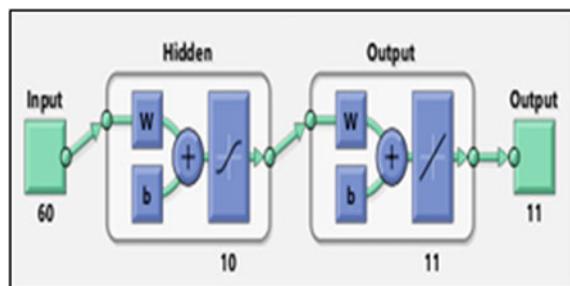
$$w_{k+1} = w_k - (J_k^T J + \mu I)^{-1} J_k e_k \quad (10)$$

where w_k = current weight, w_{k+1} = next weight, e_k = last total error, J = Jacobian matrix, μ is always positive, called combination coefficient, I = identity matrix.

The neural network configuration and internal architecture of the neural network for the faults classification are shown in Figs. 4 and 5.

Table 5 shows the number of parameters (number of inputs, number of hidden layers, number of outputs) used in the architecture of ANN.

Fig. 4 Neural network configuration for fault classification training



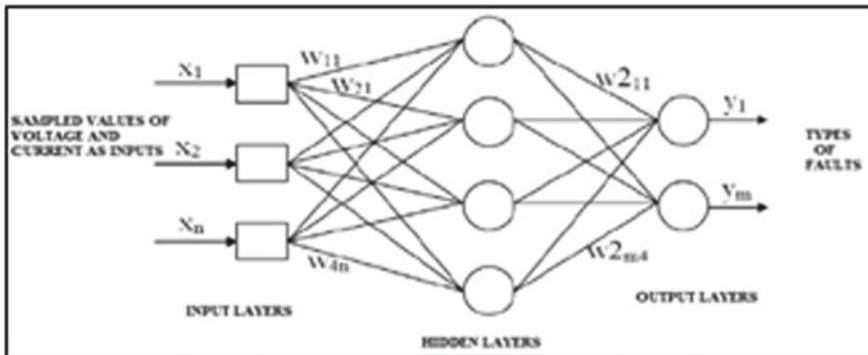


Fig. 5 Architecture of ANN

Table 5 Number of parameters for ANN architecture

Input	No. of input layers	1
	No. of input neurons	60
	No. of input patterns for fault location	60 * 209
	No. of input patterns for fault resistance	60 * 220
Hidden layer	No. of hidden layers	2
	No. of hidden neurons	10
Output	No. of output layers	1
	No. of output neurons	11
	No. of output patterns for fault location	11 * 209
	No. of output patterns for fault resistance	11 * 220

Target

No. of targets = 11 (ABCG, ABC, ABG, AB, ACG, AC, AG, BCG, BC, BG, CG)

For the variation of fault location (10 km to 190 km): No. of target patterns = 11 * 209

For the variation of fault resistance (0.01 to 95 Ω): No. of target patterns = 11 * 220

3 Results and Discussions

3.1 Simulation Results for Fault Impedance Variations

The results obtained during simulation of the IEEE-14 bus system as shown in Figs. 6 and 7 at various resistances and Figs. 8 and 9 at various locations. The given current waveform containing 20 cycles for 0.4 s. The first 10 cycles are normal and next

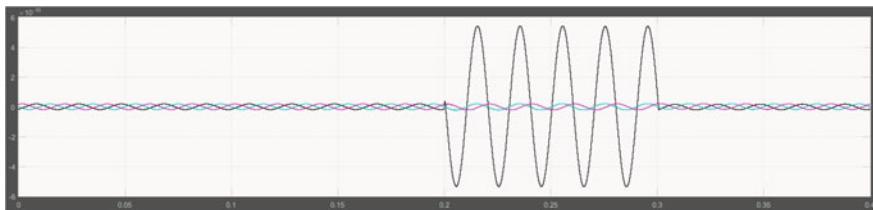


Fig. 6 Current waveforms during line to ground (AG) fault at $15\text{-}\Omega$ resistance

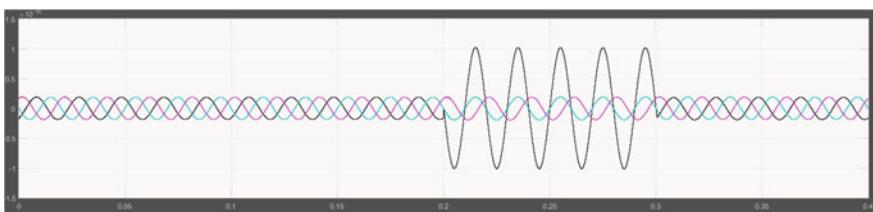


Fig. 7 Current waveforms during line to ground (AG) fault at $75\text{-}\Omega$ resistance

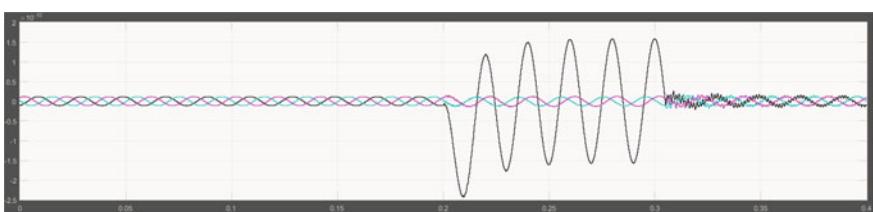


Fig. 8 Current waveforms during line to ground (AG) fault at 120 km

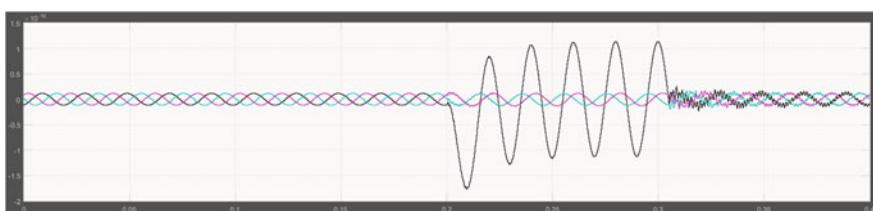


Fig. 9 Current waveforms during line to ground (AG) fault at 90 km

10 cycles are abnormal. As seen from the diagram the fault is occurring at 0.2 s containing transient and sub-transient components of current. The current of phase A sharply rises with line to ground fault at phase A and then settled to normal after

0.3 s for 15Ω resistance but the current rise for 75Ω is low as the fault impedance increases. Similarly, this form of data is collected for various resistances for all types of shunt faults.

3.2 Simulation Results for Fault Location Variations

The given current waveform contains 20 cycles for 0.4 s. The first 10 cycles are normal and next 10 cycles are abnormal. As seen from the diagram the fault occurs at 0.2 s containing transient and sub-transient components of current. The current of phase A rises sharply with the line to ground fault at phase A for 120 km than the current rise for location at 90 km is low as the location varies. Similarly, this form of data is collected for various locations for all types of shunt faults. Similarly, different voltage values are taken.

3.3 ANN Classifiers Results for Fault Impedance Variations

Figures 10 and 14 show the confusion matrix for ANN fault classifier for different fault impedances and for different fault locations. Actually, the confusion matrix gives the accuracy of ANN classifier. It will list the correct classifications for all types of shunt faults.

From Fig. 10, the overall accuracy for the classification of faults is 95% which means 95% data is correctly classified.

For fault classification, at different resistances, a neural network takes 11 epochs during the training of neural network and mean square error becomes minimum of 0.00908 as shown in Fig. 11. The error obtained during the training of ANN classifier in terms of error histogram is shown in Fig. 12 and the training state during classification of faults as shown in Fig. 13.

3.4 ANN Classifier Results for Fault Location Variations

From Fig. 14, the overall accuracy for the classification of faults is 97.1%, which means 97.1% data is correctly classified (Fig. 14).

For fault classification, at different fault locations, the neural network takes six epochs with less training time during the training of neural network and mean square error becomes minimum of 0.00698 as shown in Fig. 15. The error obtained during the training of ANN classifier in terms of error histogram is shown in Fig. 16 and the training state during classification of faults as shown in Fig. 17.

Confusion Matrix												
Output Class	1	2	3	4	5	6	7	8	9	10	11	
	19 8.6%	7 3.2%	0 0.0%	0 0.0%	0 0.0%	2 0.9%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	67.9% 32.1%
	1 0.5%	13 5.9%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	92.9% 7.1%
	0 0.0%	0 0.0%	20 9.1%	0 0.0%	0 0.0%	0 0.0%	1 0.5%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	95.2% 4.8%
	0 0.0%	0 0.0%	0 0.0%	20 9.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
	0 0.0%	0 0.0%	0 0.0%	0 0.0%	20 9.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	18 8.2%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	19 8.6%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	20 9.1%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	20 9.1%	0 0.0%	0 0.0%	100% 0.0%
	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	20 9.1%	0 0.0%	100% 0.0%
	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	20 9.1%	100% 0.0%
	95.0% 5.0%	65.0% 35.0%	100% 0.0%	100% 0.0%	100% 0.0%	90.0% 10.0%	95.0% 5.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	95.0% 5.0%

Fig. 10 Confusion matrix showing an overall accuracy of 95% for different types of faults

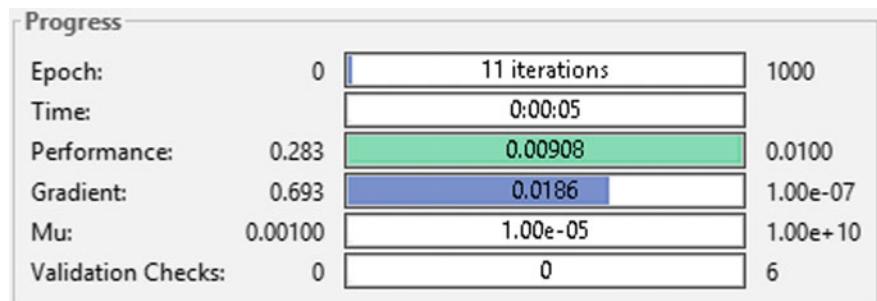


Fig. 11 Training state configuration



Fig. 12 Error histogram

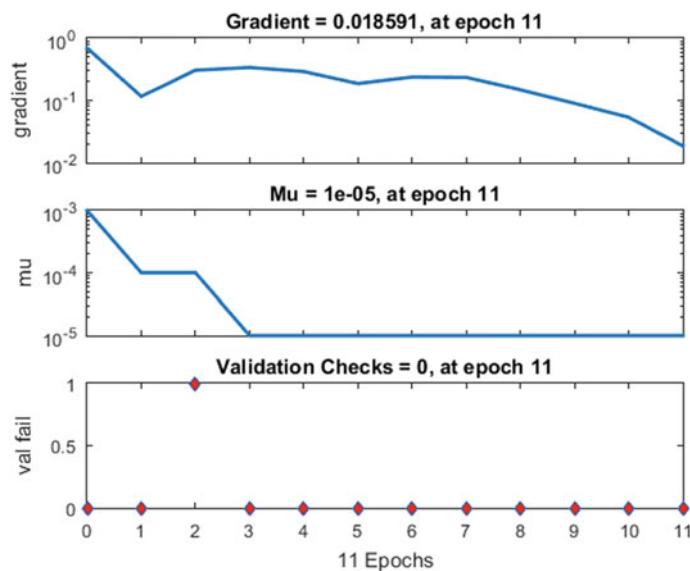


Fig. 13 Training state plot

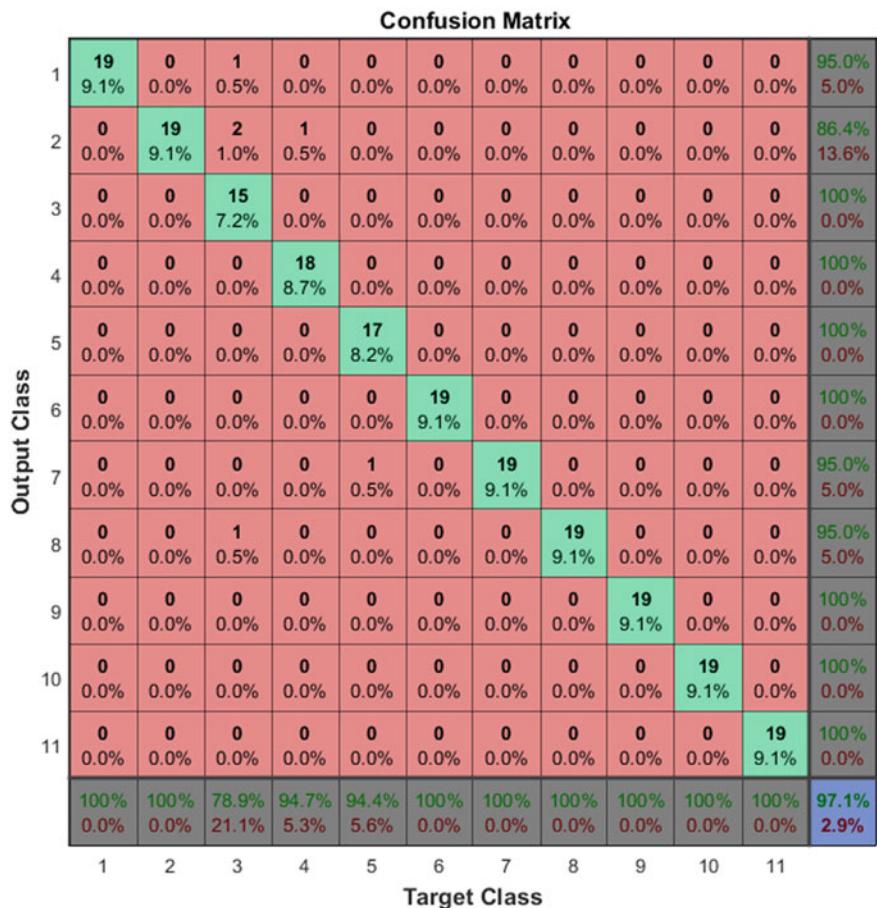


Fig. 14 Confusion matrix showing overall 97.1% accuracy for all types of faults

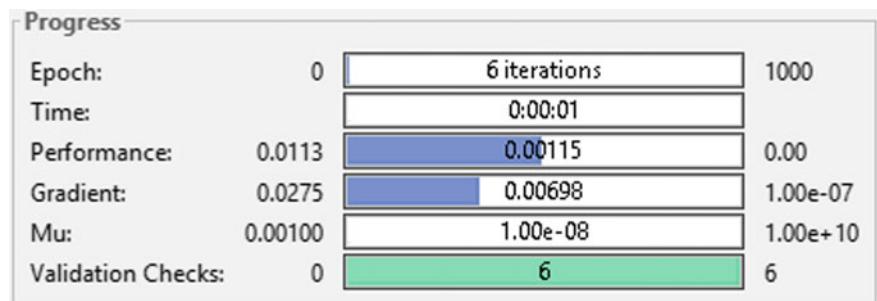
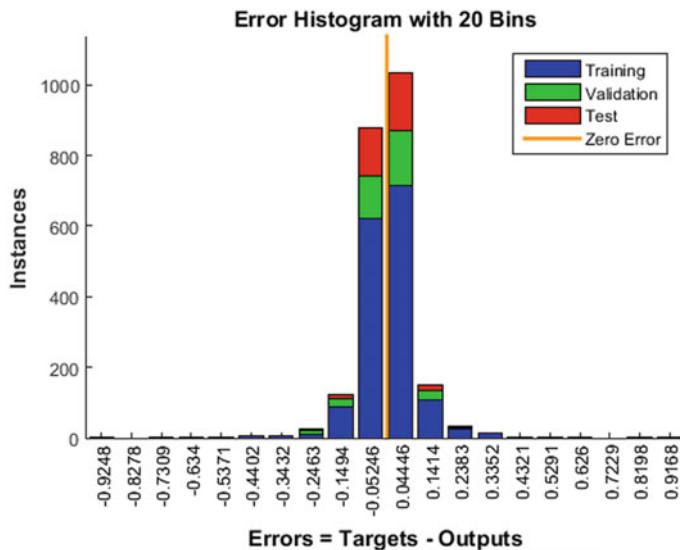
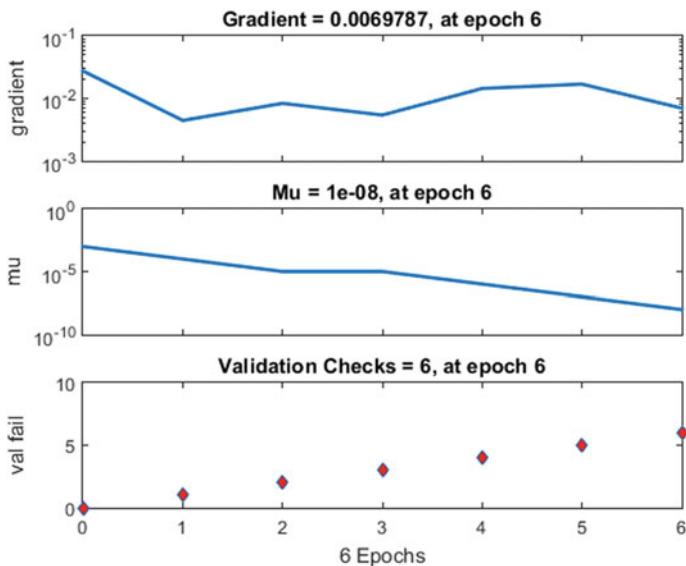


Fig. 15 Training state configuration

**Fig. 16** Error histogram**Fig. 17** Training state plot

4 Conclusion

ANN-based techniques have been used to classify the faults in the past 20 years. These algorithms depend on identifying the various patterns associated with the impedance information and learn the patterns from previous information during training time. The proposed approach uses a neural network to classify the different faults namely phase to ground, phase to phase, three-phase fault, and double line to ground faults at different fault locations and involve varying fault impedances in a transmission line of a power system with very high accuracy. The artificial neural network can classify the nonlinear relationship between measured signals by identifying different patterns of the voltage and current signals during fault conditions. As a supervised algorithm, the Levenberg–Marquardt algorithm is used for training the data. With ANN fault classifier, the overall average accuracy of 95% is obtained in case when variation in fault impedance is considered, however, it is 97.1% in case when different locations of fault are considered in IEEE 14-bus system.

References

1. O.A.S. Youssef, Fault classification based on wavelet transforms, in *IEEE/PES Transmission and Distribution Conference and Exposition*, vol. 1 (2001), pp. 531–536
2. O.A.S. Youssef, New algorithm to phase selection based on wavelet transforms, in *IEEE Transactions on Power Delivery*, vol. 17 (2002), pp. 908–914
3. M.B. Hessine, H. Jouini, Fault detection and classification approaches in transmission lines using artificial neural networks, 978-1-4799-2337-3/142014 IEEE (2014)
4. M. Oleskowicz, D.V. Coury, R.K. Aggarwal, A complete scheme for fault detection; classification and location in transmission lines using neural networks, in *Development in Power System Protection*, Conference Publication No. 479 (2001)
5. A. Nag, A. Yadav, Fault classification using artificial neural network in combined underground cable and overhead line, in *IEEE International Conference on Power Electronics. Intelligent Control and Energy Systems ICPEICES* (2017)
6. M. Jamil, in *Fault Detection and Classification in Electrical Power Transmission System Using Artificial Neural Network*, vol. 45 (Springer Plus, 2017)
7. A. Maheshwari, V. Agarwal, S.K. Sharma, Transmission line fault classification using artificial neural network based fault classifier. *Int. J. Electr. Eng. Technol. (IJEET)* **9**, 170–181 (2018)
8. J. Zhang, Z. He, S. Lin, Y. Zhang, Q. Qian, An ANFIS-based fault classification approach in power distribution system. *Int. J. Electr. Power Energy Syst.* **49**, 243–252 (2013)
9. S. Mishra, A Baral, Classification of power system faults using voltage Concordia pattern feature aided PNN, in *IEEE Conference Publications* (2016)
10. E.B.M. Tayeb, Faults detection in power systems using artificial neural network. *Am. J. Eng. Res. (AJER)* **2**, 69–75 (2013)
11. M. Jamil, A. Kalam, A. Ansari, M. Rizwan, Generalized neural network and wavelet transform based approach for fault location estimation of a transmission line. *Appl. Soft Comput.* **19**, 322–332 (2014)

12. K. Saravanababu, P. Balakrishnan, K. Sathyasekhar, Transmission line fault detection, classification, and location using discrete wavelet transform, in *International Conference on Power, Energy and Control (ICPEC)*, IEEE (2013), pp. 233–238
13. M. Saini, A.A. Mohd Zin, M.W. Mustafa, A.R. Sultan, R. Nur, Algorithm for fault location and classification on parallel transmission line using wavelet based on Clarke's transformation. Int. J. Electr. Comput. Eng. (IJECE) **8**, 699–710 (2018)

IoT Botnet: The Largest Threat to the IoT Network



Smita Dange and Madhumita Chatterjee

Abstract Adoption of the IoT technology is expanding exponentially. It is capable of providing a better service. IoT technology is successfully implemented on the bulb, refrigerator, air conditioner, washing machine, wristwatches, mobile phones, etc. Gartner report reflects that growth in the number of IoT devices is massive. By 2025, the number of IoT devices may reach up to 50 Billion. This growth poses an enormous range of challenges. The challenges are communication, interoperability, integration, data handling, privacy, and security. The major challenge is security. This paper focuses on different types of possible attacks on IoT and how the IoT botnet is gaining more attention and becoming a major attack. It highlights the key difference between traditional botnet and IoT botnet. Review of the existing techniques to deal with a botnet as well as the urge for a different technique to deal with IoT botnet is discussed.

Keywords IoT · Botnet · Botnet detection techniques

1 Introduction

A broad interpretation of the Internet of Things (IoT) is, it enables human-to-thing or thing-to-thing(s) communications [1]. Things are end devices which refer to sensors, human or any object having the potential to request/provide a service. Interconnection among things is complex as heterogeneous entities are involved [2]. IoT is implemented in every domain like agriculture, health care, food supply management, pharma supply management, environmental monitoring, and smart home. IoT has a heterogeneous environment and resource constraint devices, i.e., low memory

S. Dange (✉)
Fr. C. Rodrigues Institute of Technology, Vashi, India
e-mail: smita.dange@fcrit.ac.in

M. Chatterjee
Pillai HOC College of Engineering and Technology, Rasayani, India

and low computing power. These resource constraint devices create a hurdle for providing “one size fits all” security solution in IoT and at the same time, it increases challenges to IoT environments.

Due to the large-scale deployment, manufacturers do not consider the security of these devices. Many of the devices come with the fixed key which cannot be changed. Default username and password are the same for the devices which are manufactured in bulk. As IoT devices are resource constraint and not manufactured with built-in security principle, they are more vulnerable. Considering the growth of the IoT network, these devices are a prime security concern. The vulnerability of IoT devices open doors for different types of attacks. Botnet formation is one of the attacks which spread fast and impacts substantially. Literature describes that considerable work has been done to deal with the traditional botnet [3].

The main contribution of this paper is (1) The study of the recent major attacks on IoT system along with a listing of the possible attacks on the IoT system at the physical and network layer. (2) Overview of the IoT botnet consisting of the evolution of IoT botnet, architecture, lifecycle, and comparison between traditional botnet with IoT botnet. (3) Case study of Mirai botnet (4) Overview of the existing tools and techniques to detect botnet (5) Discussion on the exigency for prevention technique in IoT botnet.

2 Background

2.1 Recent Attacks

Different types of attacks are performed to breach security around the world from a very long time. The specialty of IoT attacks is its scale and simplicity. The enormous growth in the IoT networks creates an impact of attack at a larger scale. IoT devices are more vulnerable so can easily breach the security. It is estimated that the IoT will remain a target and attack vector for years [4]. Table 1 highlights the recent attacks.

2.2 IoT Attack Vectors

IoT attack vector span is very large. The list of all potential attacks especially occurring in the physical or network layer of the IoT architecture is mentioned in Table 2 [1, 5–7].

Table 1 Details of recent attacks in IoT

Year of attack	Infected devices/industry	The mechanism used for attack	Type of attack	Severity of attack
2018	St. Jude Medical's implantable cardiac devices	Hacker identified vulnerability present in the transmitter. Transmitter connected to a device which could read the device data. The transmitter was sharing data with a physician. Gaining control on transmitter hacker got control of the device data	Data privacy	Gaining control on healthcare-related device is dangerous as it is playing with human life
2016	Dyn internet service provider	The simplest method used by Mirai malware was to use default credential, i.e., list of the standard username and passwords. With this default credential, brute force attack is made to gain access to cameras, routers, etc., Once IoT device becomes a bot it searches for next vulnerable device	DDOS attack using IOT Botnet	Dyn suffered from DDOS attack. Twitter, the Guardian, Netflix, Reddit, and CNN websites were down
2015	SCADA system of the power grid system of Ukraine	Hacker gained control over energy distribution system and power cut happened for three–four hours in Ukraine area	DDOS attack	Ukraine area was without power supply for three–four hours

Table 2 List of attacks on IoT devices

Sr. No.	Attack name	Layer	Working	Impact
1	Jamming	Physical	Signal-to-noise ratio decreased to the level so that communication at the receiver side will be completely disturbed	Result into DOS attack
2	Sinkhole attack	Network	A sinkhole attacker's attack by offering an optimal path to reach the base station. All information passed through can be recovered by the attacker	Loss of data confidentiality
3	Node tampering	Physical	Modify the functionality/data of the node	Loss of data integrity
4	Blackhole attack	Network	Insert a new node or compromise the node in the existing network so that all the neighbors of this node will change the routing table and transmit the data from this node only. The node once received the packet will never forward it	Loss in data
5	Wormhole attack	Network	Attack has one or more malicious node and a tunnel between them. The attacking nodes capture the packets from one location and transmit them to other distant located node	The loss in data confidentiality and integrity
6	Sybil attack	Network	A single node can have multiple identities. These multiple identities can be used to spread malware and masquerade	Loss of integrity
7	Selective-forwarding attack	Network	Malicious node acts as a router and decides which packets to forward and which to drop	The loss in data privacy and integrity
8	Hello flood attack	Network	In this attack, continuous sending of message HELLO to discover neighborhood take place. Receipt of this message will be busy in replying the message. Ultimately, it will increase the congestion in the network and the device will consume its energy	It results in a DOS attack

(continued)

Table 2 (continued)

Sr. No.	Attack name	Layer	Working	Impact
9	Man-in-the middle attack	Network	Attackers secretly replay and possibly alter the communication between two parties who believe they are directly communicating with each other	Loss of confidentiality
10	DOS attack	Network	The attacker makes a machine or network unavailable to the legitimate users	Unavailability of the system
11	Flooding attack	Network	To increase the congestion, one or more malicious nodes send messages regularly	Unavailability of the system
12	Replay attack	Network	In reply attack, resending of the same data takes place	The loss in data privacy
13	Sleep deprivation attack	Physical	The main objective of the sleep deprivation attack is to keep the device awake and doesn't allow it to go in energy conservation mode	The battery will drain out fast and the node will stop working. The system will be unavailable
14	Malicious node injection	Physical	A new malicious node is placed in between two or more nodes. Hacker has control on this newly inserted node, so he can modify the data passes through this node	Loss of data integrity
15	Botnet formation	Physical and network	In this attack, machine in the network gets converted into a bot (i.e., software robot). This bot finds more vulnerable nodes and converts them into bot and forms botnet. This is a looping process. Eventually, all the machines in the network become a bot. Using botnet you can gain control on the network, data can be hacked and, DDOS attack can be performed	The system will be not available/loss of confidentiality or can result in loss of integrity

3 Overview of IoT Botnet

IoT bot represents a software robot which scans for vulnerable devices and once found converts it into bot just like a traditional bot. It is an automated process of extending malware. IoT botnet is a network of bot, i.e., infected machines. IOT botnet is controlled by botmaster who execute coordinated activities with the help of these bots. The coordinated activity could be DDOS attack, spamming, phishing campaign, click fraud, and spyware [8–12]. IoT devices turn into bot due to lack of primitive security, virus infection, or opening a malicious email attachment.

3.1 IoT Botnet Evolution

Evolution of traditional botnet and the emergence of IoT botnet is given in Table 3 [13, 14].

3.2 IoT Botnet Architecture

Traditional botnet in conjunction with IoT botnet shares the same architecture. It can be classified as a centralized botnet and P2P botnet. In both the categories, the first step is to determine the vulnerable devices in the network and acquire access over these devices. Next step is to convert these devices into bot by downloading the bot binary source from the Command and Control server. In a centralized architecture, command and control server is fixed but in decentralized architecture (i.e., P2P

Table 3 Year wise evolution of botnet

Year	Description
1988	Designed Internet's first worm "Phone home" by Robert Morris, Jr., a Cornell
1999	Sub7 and Pretty Park malware used to listen to IRC channels
2004	Phatbot malware designed to listen to P2P architecture
2006	Zeus (Zbot) malware was the first malware used to perform cyber-attack on the banking sector
2008	Grum malware had the capability to delete billion of the message in a day
2011	'Game over Zeus' was capable of dealing with a P2P protocol
2013	Security professionals report the first android botnets, such as MiscoSMS
2016	Mirai botnet used IoT devices. It is the first IoT botnets which spread on thousands of devices
2017	It is expected and predicted that IoT botnets will continue to expand looking at the growth of IoT devices. Needs a better strategy to handle IoT botnet

Table 4 Comparison between centralized botnet and P2P botnet

Botnet architecture		
	Centralized botnet	P2P botnet
Protocol used	IRC, HTTP	P2P protocol
Pros	Easy to deploy and administer	No single point of failure
Cons	Single point of failure	Coordination of bots is difficult compared with a centralized architecture
Detection	Easy	Difficult
Life cycle	<p>The life cycle has four phases</p> <ul style="list-style-type: none"> – Initial infection – Connection with command and control – Perform botnet attack – Searching for other vulnerable devices 	<p>The life cycle has four phases</p> <ul style="list-style-type: none"> – Initial infection – Connection with command and control – Perform botnet attack – Searching for other vulnerable devices

botnet) identifying CnC server location is challenging. It follows a pull/push communication mechanism. Table 4 shows a comparison between Centralized botnet and P2P botnet [8–10].

3.3 IoT Botnet Life Cycle

IoT botnet has a similar lifecycle as of traditional botnet. It consists of four phases. The first phase is known as the initial infection, followed by command and control as the second phase. Third and fourth phase deal with the attack and post-attack part [15].

3.4 Traditional Botnet Versus IoT Botnet

A traditional botnet is consisting of compromised computers or servers. It is often mentioned as zombies. Zombies are infected with malware that allows an attacker to control them, carrying out tasks on their behalf. Botnet owners or herders are able to control these infected machines in the botnet by means of a covert channel such as Internet Relay Chat (IRC) or peer-to-peer. An IoT botnet consists of compromised IoT devices, such as cameras, routers, DVRs, wearables, and other embedded technologies, infected with malware. This malware allows an attacker to control the devices, carrying out tasks similar to a traditional botnet. The comparison between the traditional botnet and IoT botnet has been given in Table 5.

Table 5 Comparison between a traditional botnet and the IoT botnet

Evolution of botnet		
	Traditional botnet	IoT botnet
Architecture	Centralized and P2P	Centralized and P2P
Life cycle	Centralized architecture has four phases—initial infection, connection with command and control, attack and post-attack	Same as traditional
Devices used	Computers, servers	Camera, smart refrigerators and smart ACs, fitness watch, health tracking devices
Detection mechanism	Easy	Difficult
Impact	Less compared with IoT botnet	Very high as the spread is fast and larger scale

Traditional botnet and IoT botnet architecture and lifecycle are similar. The major difference found among two is the underlying devices used to form the network. Characteristics of the IoT devices emphasize a different approach to deal with IoT botnet. Table 6 shows the device discrimination used in traditional network and IoT network.

Table 6 Device discrimination used in traditional network and IoT network

	Traditional network devices	IoT devices
Resource capacity	High resource devices are available	Very low processing power and less memory
Botnet detection time	Less abnormal behavior of the device can be monitor easily by user	More—IoT device has limited web-based GUI. The user interacts less with the device, so botnet goes unnoticed for a longer period
The volume of devices available for attack	Very less compared to IoT devices. Generally, in thousands	By 2020, 25 billion IoT devices are expected
In build security level	High	Very low or not deployed
Spread/impact	Less compared with IoT devices	No. of devices are very large and internet connection so spread is fast and loss is more
Availability of devices for attack	It depends upon the network	IoT devices are available 24/7. More stable source for the attack
Examples	Machines (desktops/laptops)	Smart devices include refrigerator, AC, camera, fitness watch

IoT botnet requires to be considered separately from traditional botnet when providing a solution considering the nature of the IoT devices.

4 Case Study: Mirai Botnet

On September 20, 2016 the website of respected security journalist Brian Krebs was down. It was accomplished/Performed by using Distributed Denial of Service (DDoS) attack. This was the first Mirai attack. Just 10 days after this first attack, on September 30, 2016 “Anna-Senpai” leaked the Mirai botnet source code. Then, almost t3 months later the largest DDoS attack of its kind was targeted at a popular Dynamic DNS provider, Dyn. On October 21, 2016, an attack with a magnitude of 1.2 Tbps was directed toward Dyn over the DNS port, 53. This attack was the largest IoT botnet attack that put the number of websites down.

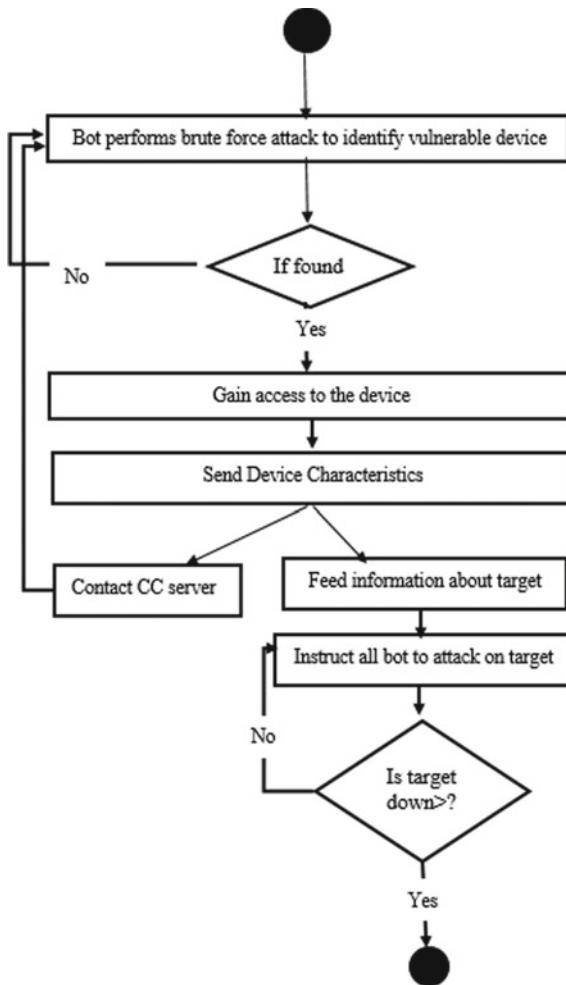
4.1 *Mirai Botnet Principle and Working*

The main goal of the Mirai botnet is to perform a DOS attack. Figure 1 depicts the working of Mirai. It can compromise IoT devices very efficiently. The command and control server runs two-socket listeners: one for Telnet connections and one for a programmatic API. The Telnet socket will listen on port 23 and route any valid connections to it to the appropriate bot or admin handler. (CnC) a portion of Mirai is written in Go, an efficient and compiled language made by Google. The API socket will listen on port 101 and route any valid attack commands sent to it to the connected bots. Each connected bot will scan the Internet for new vulnerable devices. When one is discovered, the credentials, IP address, and port used to gain access to it are sent to a loader server. This loader will output the information to the console to allow the data to be optionally stored into a file as well, and then will use the information to download and execute the malware on the device [16–18].

4.2 *Variation of Mirai Botnet*

Around 8 variations of Mirai botnet has come. Akiru, Katrina_V1, Sora, Saikin, Owari, Josho_V3, and Tokyo are few names. Since the majority of Mirai variants are copycats of the original Mirai code, they have a similar code structure.

Mirai botnet has the power to form the botnet very efficiently. After releasing the source code by Anna Senapi, researchers and hackers studied the Mirai functioning in detail. The researcher had studied Mirai in detail to provide an efficient solution. At the same time, hackers brought variation in Mirai to perform a different attack.

Fig. 1 Mirai workflow

This case study discloses that IoT botnet is a major attack on the IoT network. It demands a better solution to prevent it rather than detecting it [16, 17].

5 IoT Botnet Detection Techniques

Botnet detection techniques are broadly categorized into host-based and network-based detection techniques. A host-based technique focus on the host machine. It tracks all the activities performed/executed on the host machine, which includes processing time, access to the suspicious file, etc. It does not keep track of the network traffic. If any activity found to be suspicious, host-based intruder detection

system generates an alert and inform to the administrator. A network-based detection technique monitor network traffic. They are subcategorized into two ways, i.e., active monitoring and passive monitoring. In active monitoring, test packets are injected into the network and reactions of the network is monitored and analyzed. Based on the reactions, it can be deduced whether a human or a bot is managing the session. This method increases the traffic in the network. In passive monitoring technique, traffic is monitored and analyzes for abnormal activities. Another way of categorizing network-based botnet detecting technique is based on its underlying mechanism used for the detecting malware. The techniques are signature based, anomaly detection, DNS based, and mining based [19].

5.1 Host-Based Detection

In host-based detection technique, the individual machines are monitored. Based on certain parameters, machine behavior is analyzed and machine normal behavior is measured. In continuous monitoring process, long response time, suspected changes in the files, not able to perform a specified task, antivirus is not working, etc., is observed then it will be considered as the machine has become a bot.

5.2 Network-Based Detection

Network traffic is monitored and observed in the network-based detection technique. Different techniques can be applied to distinguish between normal traffic and malicious traffic. Protocol-level traffic can be observed based on the requirement. Malicious traffic is an indication of malware which could be the presence of a bot. Literature study exhibits network-based detection techniques are more efficient than host-based. Network-based techniques are further classified as active monitoring and passive monitoring.

Passive Monitoring

Signature-Based Botnet Detection Technique

Malware dataset is pre-request for signature-based botnet detection technique. Using the bot binaries of existing botnets, the bot's behavior can be studied. Snort and Ntop are the tools available for detecting bot based on the signature. This is a simple technique and easy to implement. The modern botnet has built up updating mechanism, botmaster update to change the signature. Multiple botnets have an identical function but different in the signature. In both cases, the signature-based detection technique fails.

Behal et al. [20]. The author proposed N-EDPS network-based detection and prevention system which will analyze the outbound traffic. The proposed system was

implemented at an educational institute. Network traffic was captured in the institute to study different types of attacks occurring in the network. Observed traffic was compared with a pre-stored malware signature. If match found, then the alert will be generated. To make the system more efficient author reduced the rule sets. Ready-made tools like BotHunter and snort are used. The proposed method is not able to detect encrypted C&C channels. It is not useful for a real-time system as it is unable to detect unknown malware.

Abbas et al. [21]. Existing signature-based detection technique, which works well with the traditional botnet. When it is applied in the IoT network to detect IoT botnet, it has found the nonfunctional overhead. This overhead cannot be handled by resource-constrained IoT devices. The proposed method is less complex. Instead of storing complete signature, the author proposed a method which stores the only a subset of signature. The author found that the proposed system gives 100% detection rate. The proposed system works in two stages. The first stage is offline where signatures are extracted from system call traces. Duplicate signatures are removed. In the second stage, these extracted signatures are used to detect malware. The author extracted benign signatures also to improve the efficiency of the system and removes the possibility of false positive detection. The proposed system works well with resource-constrained IoT devices. But it cannot detect unknown malware.

Ioulianou et al. [22] proposed a method which exploits RPL protocol (routing protocol used for low power networks including IoT networks). It monitors and measured received signal strength, packet data drop rate, packet sending rate and, no. of nodes in the IDs. The router will act as a detector. All traffic will move from the router when IoT devices want to communicate with the server. The router will check and decide traffic has come from a legitimate node or malicious node. Traffic among two IoT devices will be captured by the nearest device which acts as a detector. It can detect DOS attack, sinkhole attack, selective-forwarding attack, and Clone IDs. The author uses Cooja Simulator and performs two DDOS attacks “Hello flooding” and “Version number modify”. Method work of centralized and decentralized architecture of the botnet. The proposed System only simulated not implemented.

DNS-Based Botnet Detection Technique

It is a passive detection technique. It maintains transparency with all except both-erders. The main idea behind DNS-based approach is in order to locate C&C server bots sends DNS queries. So monitoring DNS traffic will help to detect Botnet. This is the most famous and easy technique of botnet detection. CC-based botnet using IP flux, domain flux concepts are used to hide the identity of the botmaster.

Singh et al. [23]. BOTDAD-The proposed system considers hourly DNS fingerprints generated for each host and attempts to find anomalous behavior which is quite different from normal machine behavior. The system is implemented in a campus network having more than 4000 users. DNS fingerprint generation done based on different parameters like No. of DNS request per hour, number of distinct DNS request per hour, the highest number of request for a single domain, an average number of request per minute, the highest number of request per minute, number of MX record

queries, number of PTR record queries, number of distinct DNS server queried (all based on request-based features), number of distinct TLD queried, Number of distinct SLD queried (domain-based features), and number of failed/NxDomain Queries (Response based). DNS fingerprint acts as input to the anomaly detection module which checks for any value which exceeds the threshold is labeled as bot. If there is no threshold violation, it is labeled as clean.

Anomaly-Based Botnet Detection

Chen et al. [24] proposed a method that works for the small size of the botnet. It efficiently detects abnormal IRC traffic and identifies botnet activities. The homogeneous response, PING and PPNG messages, and group activity occurred in an abnormal direction in collected traffic is considered as unique characteristics for identifying IRC-based botnet. The proposed method uses a homogeneous response and group activity. Network gateways act as a collector point for IRC traffic. Source and Destination IP, Source and destination port are the attributes extracted from network traffic. Level I correlation is achieved from observed homogeneous responses and level II correlation is achieved using group activity parameter. The author assumed traffic is not encrypted. If traffic is encrypted proposed method will not give efficient results. It works for IRC based communication only.

Sajjad et al. [25] author proposed a botnet detection method for IoT botnet based on usage, communication, and access monitoring of IoT devices. It has three main components. The first component of the proposed system focus on the device. It describes the device by keeping information about the device like its usage pattern, its communication pattern, and access policy. Monitoring is the second component of the system. It monitors and keeps a record of device usage pattern at that moment. Comparator, the third component compares current usage pattern with predefined one to find an anomaly. Router/gateway is used to implement the proposed system. The author has tested method only on Mirai malware. Only an alarm is generated when the policy is violated but no further action to mitigate it is mentioned.

Dietz et al. [26] proposed a proactive detection method to block the spreading of IoT botnet by automatically scanning the vulnerable device and isolating them from the network. The proposed method work on access router level. It has two parts: scanning and isolation. IoT devices are scanned for open ports or services, and authentication check. Once a vulnerable device is identified, change the firewall rules so communication with vulnerable IoT devices is blocked. The user will be initiated via email so he can deal with vulnerability.

Erquiaga et al. [27] discussed how to find the most relevant n/w traffic characteristics used for botnet recognition. N/w traffic attribute can be classified in two ways. First is computational resources needed to extract those attributes. Further, it can be classified as low level (i.e., from raw data i.e., IP header) and high level (traffic data). Second is the source of data, i.e., Packet, flow, and payload. The author mentioned protocol and port are not useful attributes for detecting botnet. Bpp (bytes per packet) is short as bots and botmaster sends short messages. Flow per hour, Packets per flow, flow per address are important. The payload also has important attributes channel name, joins, private message, etc.

Mining-Based Detection Techniques

McDermott et al. [28] provide a solution to the problem of detecting and making consumers situationally aware when their IoT devices are infected and form part of a botnet. The author used Deep Bidirectional Long –Short-Term Memory based Recurrent Neural Network (BLSTM-RNN), in conjunction with Word Embedding. It converts string data found in captured packets, into a format usable by the BLSTM-RNN. The model was tested on Mirai botnet malware and could detect four attack vectors of Mirai botnet with high accuracy. The proposed system does not work on prevention approaches.

Nomm and Bahsi [29] proposed high accuracy unsupervised learning model. Model reduced feature set size, which enables to decrease the required computational resources. The proposed model is common for all devices. One class, i.e., benign traffic is used to train the anomaly-based detection model. Two different data sets were used. One dataset is balanced (having the same amount of normal and compromised) and another is unbalanced (were more compromised and less normal). As the main feature, Host-IP, HOST-MAC&IP, Channel, Network Jitter and Socket were chosen and subfeatures were also limited for each category. The local outlier factor (LOF), one class SVM and isolation forest (IF) methods were applied. As generating normal traffic may not be possible for some of the IoT devices so it gives a low detection rate. Proposed method work at detection level.

Chawathe [30] proposed a method based on monitoring the network activity of IoT devices. The proposed method does not require any special access to the devices and adapts well to the addition of new devices. The author worked on the recently released dataset. The first evaluation is done on an existing dataset with all available attributes (115 for the chosen dataset) and the author finds a method to select appropriate attribute by using OneR classifier, tree-based J48, and information gain. The author again evaluated the given dataset with selective attributes and found the same accuracy. It does not discuss about prevention strategy.

Bahsi et al. [31] mentioned that minimizing the required features for classification is highly needed for overcoming scalability and computation resource problem in IoT environments. The author applied feature selection to minimize the number of features required in detecting the IoT bots and provides interpretable results with a multi-class classifier based on the shallow method decision tree. Deep learning method may give very accurate result but need high computation power and results are not easily interpretable. The author used a cross-validation technique for selecting features and trained decision tree classifier. The author does not prefer device-based classifier as it will not scale when the network grows. The output of decision tree classifier is easily interpretable and acts as input for the signature-based intrusion detection system.

Prokofiev et al. [32] developed a model of logistic regression which allows estimating the probability that a device initiating a connection is running a bot. The provided model is applicable for detection of botnets, which are propagated through brute force attack using the TELNET and/or SSH protocols. Logistic regression is a statistical model used to estimate the probability of an event based on values of a set of variables—predictors. To create the logistic regression model, the following parameters were selected as predictors: destination port, open-source ports, number of requests, even number of requests, mean interval between requests, requests on other ports. Mean size of packets, delta for packet size and, mean entropy of packets. Detection is at propagation level (botmaster identified the vulnerable device or gain access). The proposed system was only simulated not implemented.

Meidan et al. [33]. proposed a method based on deep autoencoders to detect anomalous network traffic generated from compromised IoT devices. The author evaluated this method on 9 commercial IoT devices, which were infected with Mirai and Bashlite, most widely known IoT-based botnet. IoT devices have baseline behavior and diversion in this behavior helps to detect the anomaly. So the author suggested finding the predictability of IoT devices can be quantified, finding the relationship between static and dynamic features of IoT devices, and ranking them is an open issue.

Sun et al. [34]. used a machine learning based method. Attackers in the same botnet can be put into one cluster. To find this, temporal aspect is taken into consideration and Multivariate Hawkes process, and Graph-based clustering approach is used. 10 honeypots were deployed worldwide and data is collected from it. Data collected for 1 year. If attackers perform attacks in an asynchronous manner, then the proposed system will not able to give the same result.

Mining-based techniques are getting more popular due to machine learning techniques. Table 7 gives summary of all mining based detection techniques.

Table 7 Summary of mining based detection techniques

Literature	Machine learning technique	Provides detection/prevention	Remark
McDermott et al. [28]	Deep neural network	Detection	Tested only on four attack vectors of Mirai botnet
Sven et al. [29]	Dimension reduction	Detection	Separate model for each device outperforms than the common model. Generating normal traffic for some IoT device is difficult
Sudarshan et al. [30]	Rule-based and tree-based classifier	Detection	Simple classification technique with smaller dataset gives higher accuracy
Bahsi et al. [31]	Deep learning, cross-validation for feature selection	Detection	Focus is on reducing the features need to detect the malware. Deep learning requires high computation power
Prokofiev et al. [32]	Logistic regression	Detection	It detects the botnet which is propagated through brute force attack using the TELNET and/or SSH protocols. The proposed system is only simulated
Meidan et al. [33]	Deep learning autoencoder	Detection	Each IoT device has limited functionality. So it is easy to predict the behavior of IoT device
Sun et al. [34]	Graph-based clustering	Detection	Assumes attack occurs in a synchronous manner only

6 Discussion and Analysis of Different Detection Techniques

Based on the study of existing techniques, the anatomy of botnet detection techniques is given in Fig. 2.

The network-based detection technique is more preferable than host-based techniques. The success rate of detecting a bot in the network-based the approach is more than host-based in a traditional network. In IoT network also, the network-based technique would be preferred considering the resource-constrained nature and the volume of the IoT devices. Signature-based is the simplest method to implement. Readymade tools are available based on the signature-based method. The success rate of the signature-based method depends upon the malware database. If the malware database is covering all the existing malware signature and updated frequently then the signature-based method gives 100% accuracy. In DNS based method major concern is to detect the location of the CnC server and then detach it from the IoT network. IP flux, domain flux techniques make it difficult to locate the CnC. This approach works well with centralized architecture only. Anomaly-based helps to detect known and unknown malware. In honeypot data collection is the first phase of the system and analysis of the collected data is part of the second phase. So, the honeypot is a good option for detection but cannot be used for prevention mechanism. Mining-based methods using machine learning are becoming more popular. Analyzing traffic and applying classification/clustering algorithm on that to detect the malicious activity helps to detect the bot. The approach works well in centralized and decentralized architecture. Existing methods work well on detection mechanism only not focus on prevention mechanism. Table 8 gives comparison between network-based botnet detection techniques.

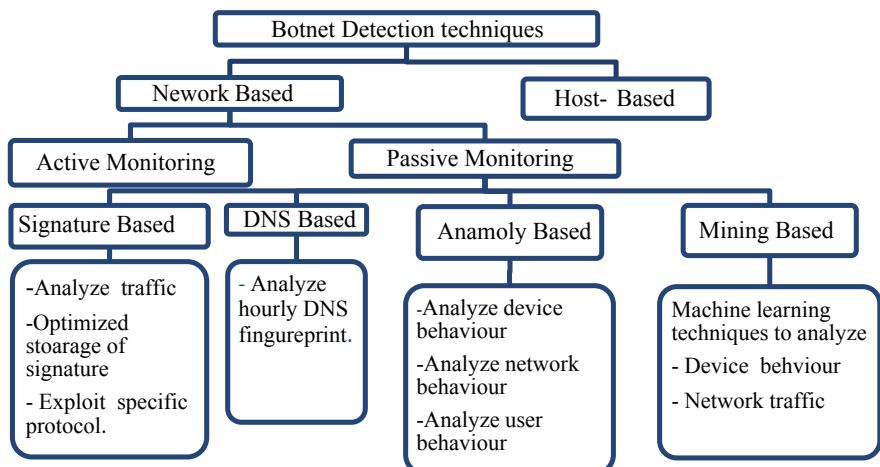


Fig. 2 An anatomy of IoT botnet detection techniques

Table 8 Comparison among network-based detection techniques

	Network based			
	Signature based	Anomaly based	DNS based	Mining based
Basic methodology/principle	Compares signature with malware dataset. If match found in the dataset it will block	Find flaws in the network like high amount of network latency, sudden increase in network traffic, presence of data traffic on unusual ports	Locate C&C server and isolate it	Sufficient amount of data is obtained from the network log file and analysis is carried out
Pros	Easy to implement. Simple approach	Works well with smaller IoT network	Easy to implement	Accuracy and efficiency is high
Cons	Efficiency depends upon the malware dataset. Not able to detect when bot master update and change the signature	Need to store baseline behavior of the user, network and, devices	IP flux, domain flux technique decreases the efficiency of DNS based detection technique	Some technique needs high computation resource
Architecture support	Suitable for centralized and decentralized botnet architecture	Suitable for centralized and decentralized botnet architecture	Supports only centralized architecture	Suitable for centralized and decentralized botnet architecture
Detects known/unknown malware	Only known	Both	Both	Both
Detect encrypted bots?	Yes	Yes	Yes	Yes
Use of tools	Yes. Snort and Ntop	Yes, BotSniffer	No	No
Mode of working	Passive monitoring	Passive monitoring	Passive monitoring	Passive monitoring

7 Open Research: Need for Prevention Mechanism to Deal with IoT Botnet

Gartner report quoted that the no. of IoT devices will reach 21 billion by 2020. IoT devices are inadequate with a built-in security feature. Recent attacks reveal that the IoT device vulnerability can be captured easily so can be easily exploited. Mirai botnet was the famous attack take place in 2016 where IoT camera and routers were

converted into the bot. Mirai used a simple technique to hack the IoT device. It used 62 possible username–password combination and performed a brute force attack to gain access to the IoT device.

Literature study highlights existing methods, which provides a solution to deal with the traditional botnet. These solutions had not considered the nature of the IoT devices. Few researchers worked on the IoT botnet but could provide solution only at the detection level [35, 36]. To handle the IoT botnet efficiently a long term strategic solution is required. Providing a solution for preventing the formation of a bot is a much more efficient solution than detection. The solution should work on centralized and decentralized architecture. Each detection technique considered a particular aspect while providing a solution. To gain a high level of success rate and provide a solution for prevention, different detection techniques can be merged with necessary modification. Considering the life cycle of IoT botnet, prevention mechanism should not allow the IoT device to enter into stage 2, i.e., communication with Command and Control Server.

8 Conclusion

Undoubtedly, all the security-related issues in the IoT need to be handled efficiently to make IoT vision as a reality. IoT-based application should provide trust, security, and privacy features on a priority basis. Out of all the potential attacks on the IoT system, IoT botnet is gaining more popularity. IoT botnet spread fast considering the IoT network as well as creates more impact as compared to other attacks. This survey has covered all the details of IoT botnet from definition to detection techniques. It has been observed that IoT botnet requires a different mechanism to deal with it. Prevention mechanism can be considered as the long-term best solution. Out of the existing techniques, network-based techniques are more efficient. Developing a new hybrid approach based on the network-based botnet detection technique to specially deal with IoT botnet would help to secure the IoT network from IoT botnet attack.

References

1. D. Singh, G. Tripathi, A.J. Jara, A survey of internet-of-things: future vision, architecture, challenges and services, in *Proceedings of IEEE World Forum Internet Things* (2014), pp. 287–292
2. L. Atzori, A. Iera, G. Morabito, The internet of things: a survey. *Comput. Netw.* **54**, 2787–2805 (2010)
3. S.A. Kumar, T. Vealey, H. Srivastava, Security in internet of things: challenges, solutions and future directions, in *49th Hawaii International Conference on System Sciences* (2016)
4. M. Binti, M. Noor, W.H. Hassan, Current research on the internet of things (IoT) security: a survey. *Comput. Netw.* **148**, 283–294 (2019)
5. L. Atzori, A. Iera, G. Morabito, The internet of things: a survey. *Comput. Netw.* **54**(15), 2787–2805 (2010)

6. J. Deogirikar, A. Vidhate, Security attacks in IoT: a survey, in *Proceedings of IEEE International Conference on I-SMAC* (IoT in Social, Mobile, Analytics and Cloud) (2017), pp. 33–37
7. S. Benzarti, B. Triki, O. Korbaa, A survey on attacks in internet of things, in *Proceedings of IEEE International Conference on Engineering & MIS*, Tunisia (2017)
8. R. Khan, S.U. Khan, R. Zaheer, S. Khan, Future internet: the internet of things architecture, possible applications and key challenges, in *Proceedings of IEEE 10th International Conference on Frontiers of Information Technology* (2012), pp. 257–260
9. O. Vermesan, P. Friess, A. Furness, in *The Internet of Things 2012* (New Horizons, 2012). [Online] Available http://www.internet-of-things-research.eu/pdf/IERC_Cluster_Book_2014_Ch.3_SRIA_WEB.pdf
10. E. Fernandes, A. Rahmati, K. Eykholt, A. Prakash, Internet of things security research: a rehash of old ideas or new intellectual challenges? *J. Syst. Attacks Deference J.* (Co-published by the IEEE Computer and Reliability Societies) (2017)
11. S.S.C. Silva, R.M.P. Silva, R.C.G. Pinto, R.M. Salles, Botnets: a survey. *Comput. Netw.* **57**, 378–403 (2013)
12. N. Mims, in *The Botnet Problem* (Elsevier, 2017)
13. K. Sha, W. Wei, T.A. Yang, Z. Wang, W. Shi, On security challenges and open issues in internet of things. *Futur. Gener. Comput. Syst.* (Elsevier, 2018)
14. Botnet Evolution Infographics, *CYRN Cyber Threat Report* (2017)
15. N. Hachem, Y.B. Mustapha, G.G. Granadillo, H. Debar, Botnets: lifecycle and taxonomy, in *IEEE* (2011)
16. A.R. Sfar, E. Natalizio, Y. Challal, Z. Chtourou, A roadmap for security challenges in the internet of things. *Digit. Commun. Netw.* (Elsevier, 2018)
17. D.E. Kouicem, A. Bouabdallah, H. Lakhlef, Internet of things security: a top-down survey. *Comput. Netw. J.* (Elsevier, 2018)
18. C. Kolias, G. Kambourakis, A. Stavrou, J. Voas, in *DDoS in the IoT: Mirai and Other Botnets* (CyberTrust by IEEE Computer Society, 2017)
19. N. Goodman, A survey of advances in Botnet technologies, arxiv (2017)
20. S. Behal, A. Singh, K. Kumar, Signature based botnet detection and prevention, in *IEEE* (2010)
21. M.F.B. Abbas, T. Srikanthan, in *Low-Complexity Signature-Based Malware Detection for IoT Devices* (Springer, 2017)
22. P.P. Ioulianou, V.G. Vassilakis, I.D. Moscholios, M.D. Logothetis, A signature-based intrusion detection system for internet of things (2018)
23. M. Singh, M. Singh, S. Kaur, in *Detecting Bot-Infected Machine Using DNS Fingerprinting* (Elsevier's Digital Investigation, 2019)
24. C.-M. Chen, H.-C. Lin, Detecting botnet by anomalous traffic. *Inf. Secur. Appl.* (Elsevier, 2014)
25. S.M. Sajjad, M. Yousaf, UCAM: usage, communication and access monitoring based detection system for IoT botnets, in *IEEE Conference* (2018)
26. C. Dietz, R.L. Castro, J. Steinberger, C. Wilczak, M. Antzek, A. Spreotto, A. Pras, IoT-botnet detection and isolation by access routers (2017)
27. M.J. Erquiaga, C. Catania, C.G. Garino, An analysis of network traffic characteristics for Botnet detection, in *CICAC* (2015)
28. C.D. McDermott, A.V. Petrovski, F. Majdani, Towards situational awareness of botnet activity in the internet of things, in *International Conference on Cyber Situational Awareness, Data Analytics and Assessment (Cyber SA)*, IEEE, UK (2018)
29. S. Nomm, H. Bahsi, Unsupervised anomaly based botnet detection in IoT networks, in *International Conference on Machine Learning and Applications*, IEEE (2017), pp. 1048–1052
30. S.S. Chawathe, Monitoring IoT networks for botnet activity, in *IEEE Conference* (2018)
31. H. Bahsi, S. Nomm, F.B. La Torre, Dimensionality reduction for machine learning based IoT botnet detection, in *International Conference on Control, Automation, Robotics and Vision (ICARCV)*, IEEE, Singapore (2018), pp. 1857–1862
32. A.O. Prokofiev, Y.S. Smirnova, V.A. Surov, A method to detect internet of things botnets, in *IEEE Conference* (2018)

33. Y. Meidan, M. Bohadana, Y. Mathov, Y. Mirsky, D. Breitenbacher, A. Shabtai, Y. Elovici, N-BaIoT: network-based detection of IoT Botnet attacks using deep autoencoders. *IEEE Pervasive Comput.* **13**(9) (2018)
34. P. Sun, J. Li, M. Zakirul, L. Wang, B. Li, Modelling and clustering attackers activities in IoT through machine learning techniques. *Inf. Sci. J.* (Elsevier, 2018)
35. E. Bertino, N. Islam, in *Botnets and Internet of Things Security* (CyberTrust by IEEE Computer Society, 2017)
36. Y. Ji, L. Yao, S. Liu, H. Yao, Q. Ye, R. Wang, The study on the botnet and its prevention policies in the internet of things, in *IEEE Conference* (2018)

Building a Trustworthy Ethical Approach to Cloud Computing



Ankita Sharma and Hema Banati

Abstract Cloud computing has gained wide acceptance by business enterprises all over the globe. With the advancement in the popularity of cloud computing, research is being directed to study the effect of various aspects affecting it. One of the most important issues that impact cloud computing is the human perception of it. The inclination of users, across the globe, towards cloud computing is affected by the amount of trust they repose in the services and the perceived level of ethics being observed in maintaining these services. This paper studies the impact of ethics and trust on Cloud computing and its subsequent effect on e-commerce. A model is proposed to incorporate trust and ethics for an improved e-commerce environment (ETCM), Ethics Trust Cloud model. Further a survey is conducted to deduce the benefits of cloud computing on e-commerce where the respondent provides opinion about the prominent factors which contribute towards building of trust in the service provider and the possible relationship between observance of ethical practices and trust-building. The data is collected by a specific questionnaire, which is designed to understand the user's perception about these concepts and their possible influence on cloud computing environment. Findings of the survey and subsequent analysis reveal a positive correlation and regression between trust and ethics.

Keywords Cloud · Ecommerce · Trust · Ethics · Regression

1 Introduction

Cloud computing and e-business are two trendy expressions these days, the prominence can be credited basically to the factor of cost adequacy. Cloud computing helps in sparing association cost of IT foundation. It helps in use of the system assets in

A. Sharma (✉)
Jagannath University, Jaipur, Rajasthan, India
e-mail: ankita.sharma@jimsindia.org

H. Banati
Department of Computer Science, Dyal Singh College, University of Delhi, New Delhi, India
e-mail: banatihema@hotmail.com

a cost-proficient and powerful way and diminishes the misfortunes of the association, assuming any, brought about by disappointment of the PC hardware like loss of information and so forth. Cloud computing empowers clients to utilize the product and equipment, without fretting over foundation subtleties and empowering center around the center administrations and assets required [1].

Cloud applications are crucial to the core of business operations of the consumers, The SLA (Service-level agreement) between providers and consumers [2] provides the much needed assurance to an organization. Cloud computing provides Trust and Security, mobility and the prospects of global expansion to an e-commerce environment. Investments get tailored to the needs of E-commerce and the associated benefits of scalability make this a very viable option for an e-commerce environment Hayes [3] discusses the trend of moving software applications on the cloud and the related trust privacy, security, and reliability challenges. Nowadays number of commercial organizations are getting inclined to use the e-commerce services and products. In May of 2011, the first public Cloud service platform was built in Hangzhou. This platform is called West Lake Cloud Computing Public Service Platform which aims to serve Chinese e-commerce industry. Five enterprises from Beijing, Shanghai and Hangzhou are the first clients of this platform [4]. Other software companies such as Ufida [5], Kingdee [6] and eAbax [7], have promoted their whole journey e-commerce and built service platforms in China. Companies have started using cloud computing for business innovation, i.e., to take the first place between the competitors. Cloud computing is also a way for companies to incorporate applications for mobile technology. Cloud computing plays an important role offering infrastructure needed for new applications and it also provides a way for organizations to incorporate the mobile applications with the system that already exists [8].

From the research conducted by IBM in 2014, we can observe that cloud computing is helping the organizations to grow. A research conducted by IBM and Economist Intelligent reveals that 67% of the companies that have dedicated a fixed amount of revenue generation for integration with cloud computing, cloud computing is helping their business to grow [8].

1.1 Benefits for Cloud Providers

Cloud computing provides incentives to the e-commerce organizations [4]

1.1.1 Make Money

This is one of the major reasons for organizations to make use of Cloud computing. In research conducted by Greenberg et al. [9] it was observed that by converting everything into a monthly charge helps the cloud providers to make profits.

1.1.2 Hold the Existing Investment

The cloud providers must make use of the existing infrastructure to enable companies to generate revenues. For example, Amazon and Google extended their private clouds and made its use public.

1.1.3 Constructing a Franchise

A large group of enterprises has begun to use cloud computing for their business, so if franchises are established the vendors will be motivated to provide Cloud to others as well.

1.1.4 Building Customer Relationship

The companies build customer relationship with the help of their services which includes Cloud services.

1.1.5 Become a Platform

Nowadays, the organizations are providing cloud services to become a platform. The infrastructure provider of Facebook plug-in application is a Cloud provider called Joyent [10]. These providers include IBM, Google, Amazon, NetSuite, Rackspace, Soft layer, telemarket, etc.

1.2 Benefits for Cloud Consumers

Here, Cloud consumers refer to companies and organizations that adopt Cloud Computing. Armbrust et al. gave reasons to explain why these organizations want to move to the cloud [11].

1.2.1 Pay as Use

Small-scale organizations do not need to invest in the initial IT infrastructure. They can simply pay as they use.

1.2.2 Reduce Operation Cost

In utility computing there is the use of virtual machine instead of physical machine, therefore the work of hardware operation is shifted from cloud consumers to cloud providers. There are additional benefits for cloud consumers for instance, start-ups can skip the hardware procurement and capital expenditure phase.

According to the Management expert, Peter Drucker “management is the responsibility for execution”. [12]. The implication of his words was that as a manager you’ll be judged on at least one thing on the extent to which you accomplish your unit’s goals. Therefore, every organization follows a technique to enhance the performance of the business.

1.3 Benefits to the End Users

The end users of Cloud service are the most important ones. They are very much like the enterprises and organizations and they require easy to use interfaces with reliability, timely delivery etc. [13].

The online marketplace is thriving, and e-commerce battle lines are being drawn between popular names such as Amazon and Alibaba [14]. Amazon is the current leader among e-commerce transactions these days. There is another dominant e-commerce company Alibaba with a market share of more than 80% in China.

Amazon is involved in B2C direct sales (business to consumer) via Amazon Web Services (AWS), and a platform for retailers to sell to buyers [15]. Alibaba, on the other hand, is basically a B2B (business to business) platform, which helps in facilitation of the sale of goods and services between businesses. However, the Alibaba Group comprises of affiliates such as Taobao Marketplace, a C2C (consumer-to-consumer) platform which is similar to eBay, and Taobao Mall (or Tmall), a B2C platform similar to Amazon.

The 2015 market share had Alibaba leading with 80% of the Chinese online marketplace, while Amazon had 60 percent of the US online sales [15]

The **Base Layer (IaaS)** of E-Commerce Cloud is shared by the infrastructure resources which connect the various service provider’s huge system and pools them together to provide services. Cloud computing allows to access data resource in secure and scalable way and allows to share the hardware resource and make use of hardware layer to run in the most likely way. The virtualization technology can be used to separate the physical hardware from the operating system. The **Platform Layer (PaaS)** of E-Commerce Cloud includes the task which earlier had been difficult to complete but now with the help of powerful hardware it is possible to complete it now: task of data storage carried out by platform layer, computation and software development, task of computation of original mass storage can be achieved, business intelligence processing possible and so on. Now choosing devices by the users and based on complexity of dealing with content the number of devices depends [16]. Also, authentication of both the user and services is considered as a specific issue

in trust and security of Cloud computing [17]. The **Application Layer (SaaS)** of E-Commerce Cloud provides the application software or services and use the e-commerce system to pay for getting the benefit of lower cost and remove wastage and make able to use more resources which help to run the business activities smoothly [18]. Cost is determined on demand-access.

This paper studies the relationship between cloud computing and e-commerce. The focus of the work is to assess the impact of ethics and trust on cloud computing. The paper proposes a distinct method to measure the practices being observed in the cloud environment. A model to integrate the trust and ethics factor in the cloud environment; is also proposed. The organization of the paper is as follows: The following section (Sect. 2) proposes a model to incorporate trust and ethics for an e-commerce organization using cloud computing and Sect. 3 describes the instrument used to assess the user's perception of trust and ethics in the cloud-based scenario, and a detailed analysis of the user feedback is also done in this section. Section 5 concludes the paper.

2 Cloud Computing Model Incorporating Ethics and Trust on E-commerce (ETCM)

Based on the study conducted and the correlation between Ethics and Trust a need was felt to incorporate them in a strategic management process (Fig. 1).

Figure 2 represents the strategic management process that is generally followed by any organization to improve the performance of the business.

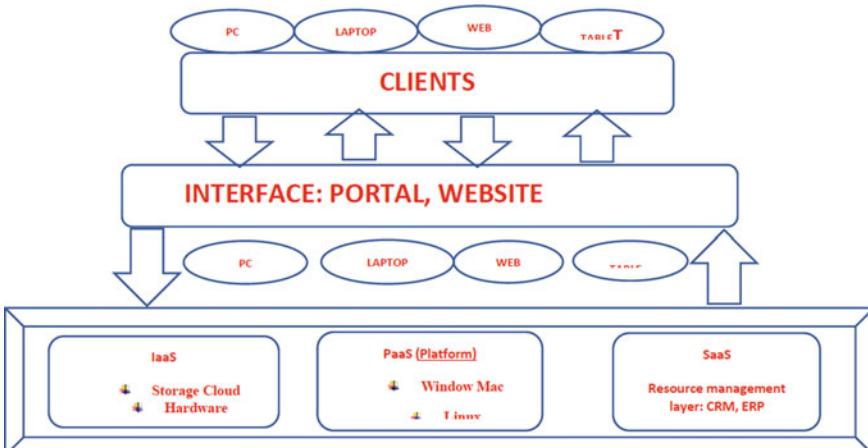


Fig. 1 An interpretation of how various services of cloud computing can be utilized in an e-commerce platform

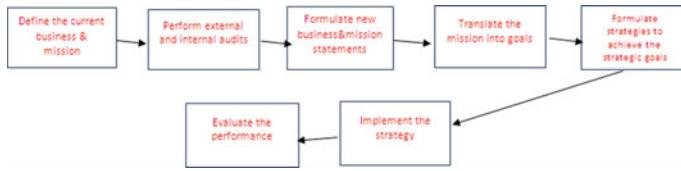


Fig. 2 Strategic management process [12]

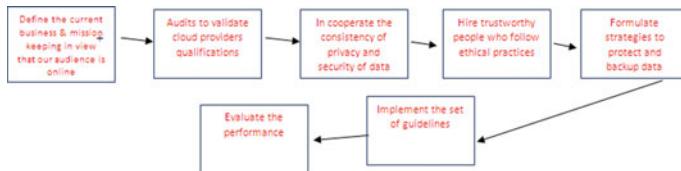


Fig. 3 Strategic management process for cloud computing

Figure 3 represents the management process following cloud computing including trust and ethics as the basic factors to run the business and industry. Cloud computing now has become an essential part of E-commerce as it deals with all the information stored online and how it can be accessed and shared. There are many benefits to the E-commerce industry with the help of which cloud computing is considered being essential, cost is taken to be the most important. Due to increasing number of suppliers of cloud services and attractiveness of investing in cloud computing it seems to be a matter of time before cloud computing takes over entire E-commerce market. Further with the analysis of the questionnaire, we concluded that cloud computing is being increasingly used in the industry and more than 50% respondents have been using cloud for more than 3 years.

Also, more than 80% of the respondents have agreed that we can trust the cloud provider to store our sensitive, personal and financial data. Ethically more than 80% of the respondents agree that the cloud providers are securing the data following certain protocols and audits.

An organization normally conducts an audit procedure to build the code of ethics to make their businesses run in an ethical manner leading to the development of trust.

2.1 *Organizational Ethics Practices*

An organization follows the code of conduct to run the business ethically. It explains that an organization should work with integrity and transparency in everything we do in accordance with our unique culture and values. Values are also influenced by the principle of trusteeship. Trusteeship also includes the cooperative commitment to utilizing the natural resources in a sustainable way and improve the communities in which we live and work.

2.2 Regular Ethical Audits to Verify the Ethical Practices

2.2.1 Begin with a Strong Foundation

An ethical audit is a comparison between actual employee behavior and the guidance for the employee behavior mentioned in policies and procedures for comparison.

2.2.2 Develop Metrics

Ethical audits might not just be black and white as financial or operational audits, but instead, they can run smoothly when ethics measures are in place.

2.2.3 Create a Cross-Functional Team of Members

A mix of professionals helps to make ethical audits more effective. Including an HR professional with an ethics and compliance manager along with the internal auditor and legal managers.

2.2.4 Keeping an Eye on Other Issues as Well

If the organization keeps an eye on issues as well it benefits the organization in other aspects as well, i.e., ethical issues in sales area may have some revenue recognition implication from a financial reporting perspective.

2.2.5 Responding Consistently and Strong Communication

A complete record of the ethical violation should be traced, and an effective procedure must be followed to run the code of ethics.

Together all of these have a positive impact on E-commerce business and industry [19].

3 To Deduce Impact of Cloud Computing to the E-commerce Sector

A survey was conducted to understand the opinions IT professionals have about ethics and trust in cloud computing. To determine the sample for the study, purposive sampling was used. Purposive sampling is a non-probability sampling technique which is used when the sample is chosen based on the researcher's judgment. This

sampling technique is also known as judgment or selective sampling. In this study, only those respondents were chosen who are working in an IT field and have working knowledge of cloud. Total 150 professionals were addressed and invited to fill the survey.

The primary study was conducted by using the questionnaire technique. A questionnaire with structured questions was developed to analyze the opinions of IT professionals who are working on cloud computing. The questionnaire was mainly divided into 3 parts. The first part was related to the demographic information of the respondents. Second part was mainly related to trust. Respondents were asked to answer the questions at nominal scale (yes/no). The questions in third part of the questionnaire were related to ethics in cloud computing and benefits of cloud computing to e-commerce industry. These questions belong to ordinal scale. Respondents are asked to give their responses on 5-point scale ranging from strongly disagree to strongly agree.

Our basic objectives of the questionnaire are

1. Which service of the cloud is being used?
2. Determination of factors for trust building in Cloud?
3. How secure is the cloud provider?
4. How are ethics and trust in Cloud related?
5. Impact of ethics and trust on Cloud computing for E-commerce.

3.1 Analysis of the Data Collected

The data analysis is done with the help of statistical software SPSS 20.0. To test the reliability of our instrument of data collection, i.e., questionnaire, Cronbach Alpha test is used. Cronbach Alpha test of reliability describes how much consistency is present in the responses to the questionnaire [Cronbach alpha is a measure that is used to access the reliability or internal consistency of a set of scale or test item]. A value greater than 0.7 considered to be good. In our responses, value of Cronbach alpha is 0.842 which shows a good consistency of responses (Table 1).

The detailed analysis of the questionnaire is done with the help of frequency tables and bar diagrams. The first part of questionnaire is related to demographic profile of the respondents. It includes questions like their designation, in which industry they are working, how much experience they have, etc. The analysis is shown in following tables and diagrams. The objectives mentioned above are analyzed with the questionnaire below.

Table 1 Reliability statistics

Cronbach's alpha	N of Items
0.842	18

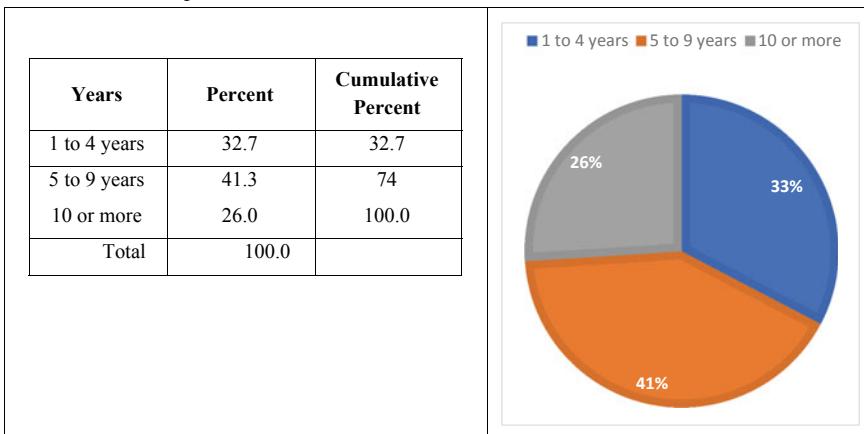
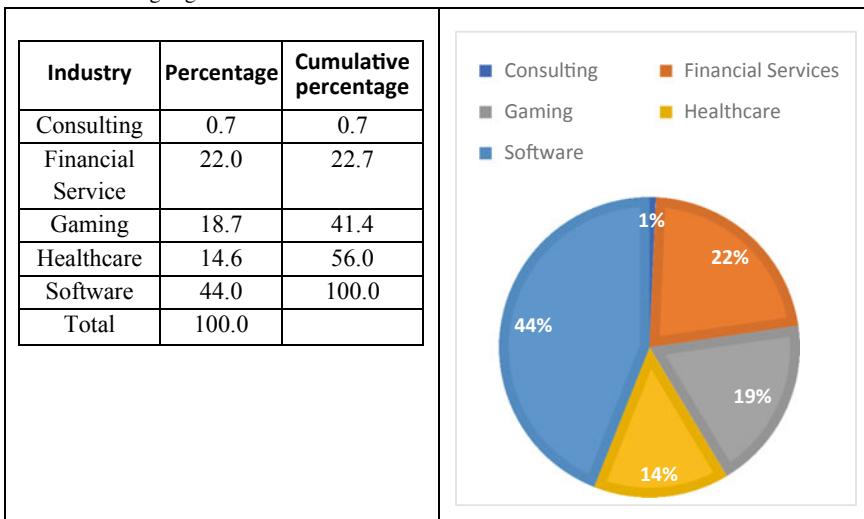
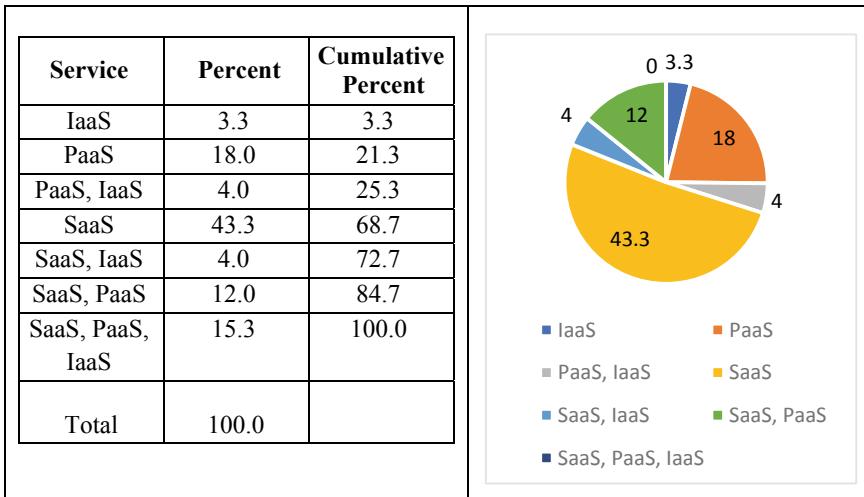
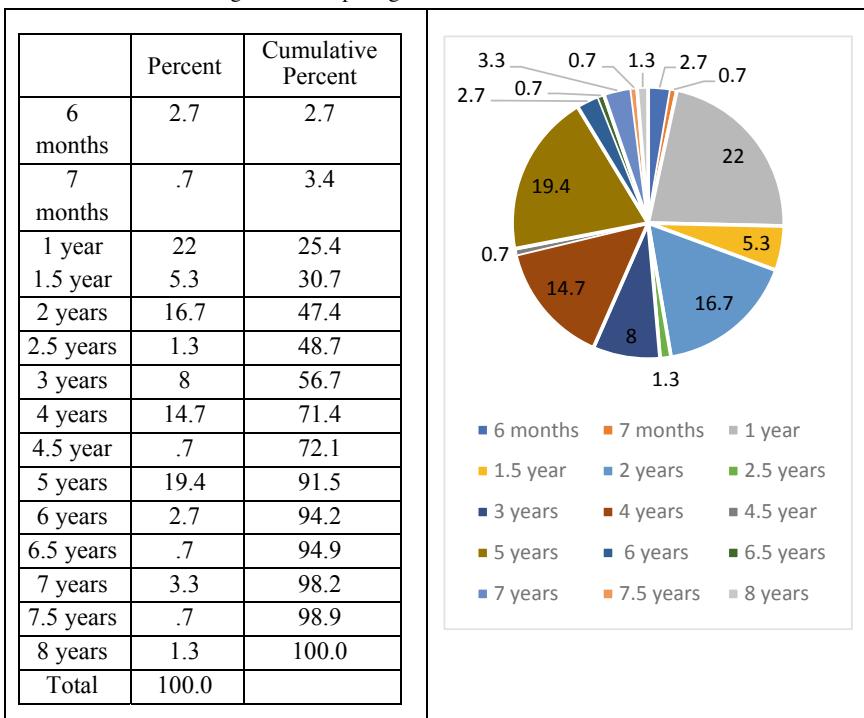
PART A**Q1. Years of experience (Table 2).****Table 2** Years of experience**Q2. Which of the following best describes sector (industry) of your organization? (Table 3)****Table 3** Working organization**PART B****Q1. What sort of Cloud Service have you used? (Table 4)**

Table 4 The Cloud service

Q2. Since how long you have been using Cloud Computing? (Table 5)

Table 5 Duration of using cloud computing

Q3 (i). Do you feel the way/place the data is sorted is of relevance to you?

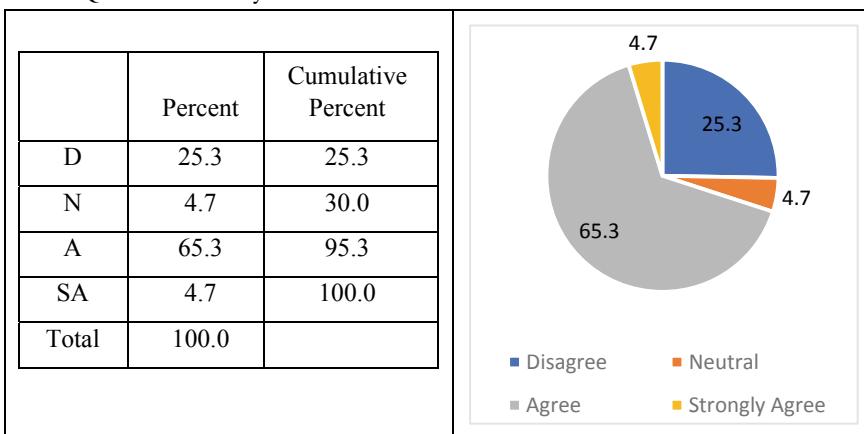
Disagree—D

Neutral—N

Agree—A

Strongly agree—SA (Table 6).

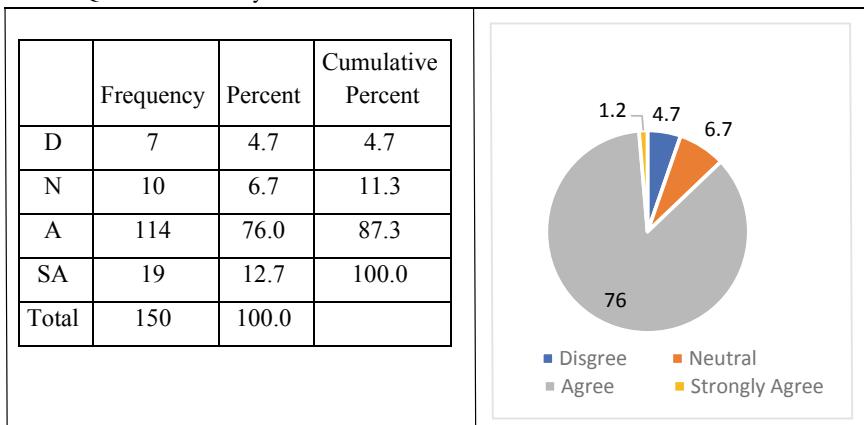
Table 6 Questionnaire analysis



It is evident from the above table and graph that majority of respondents agree that the place/way the data is sorted is of relevance to them. One-fourth of respondents disagree to this statement.

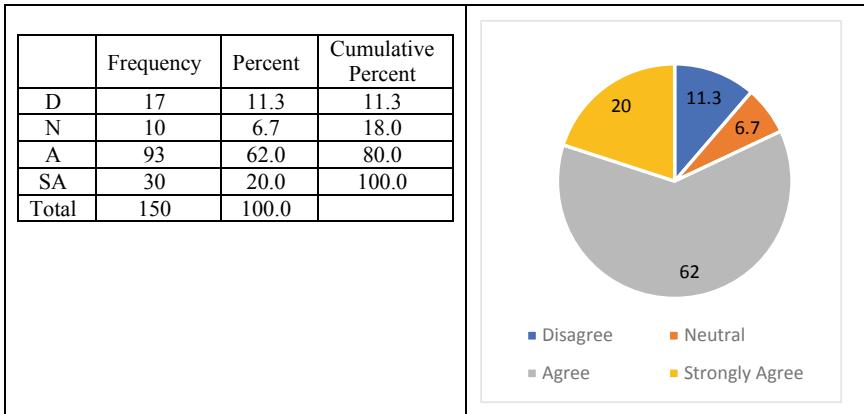
(ii). Do you feel there is a need for jurisdiction corresponding to your data in consistency of privacy and security aspects provided by the provider? (Table 7)

Table 7 Questionnaire analysis



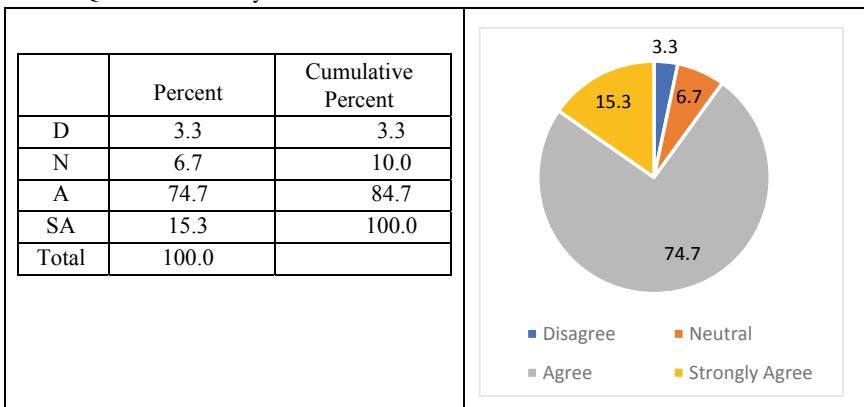
(iii). Has the provider provided storage as and when required? (Table 8)

Table 8 Questionnaire analysis

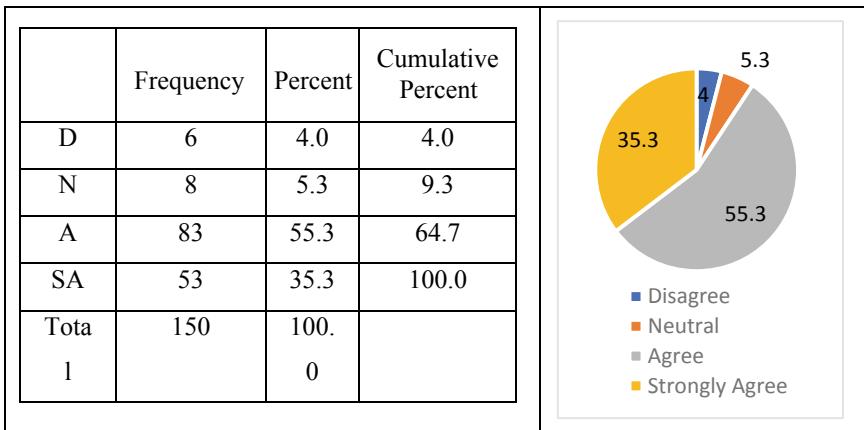


(v). Do you have the assurance by the cloud provider that in case of an unfortunate instance it breaks down or gets acquired by other company then your data will still be available? (Table 9)

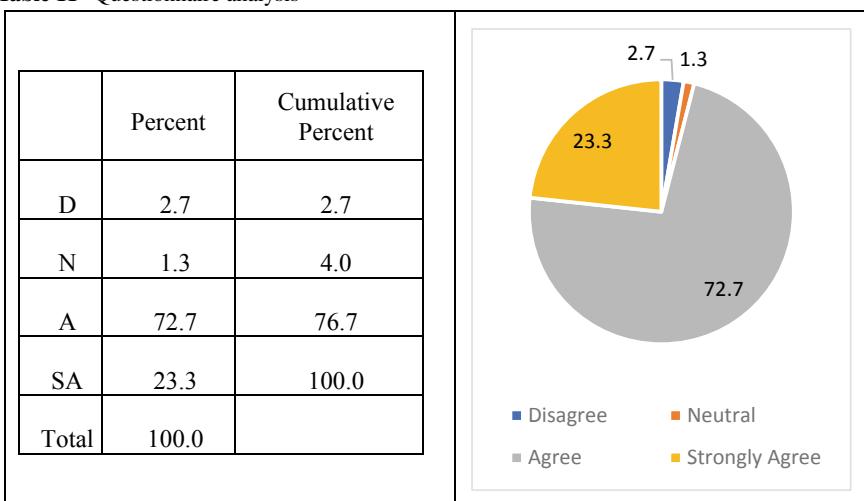
Table 9 Questionnaire analysis



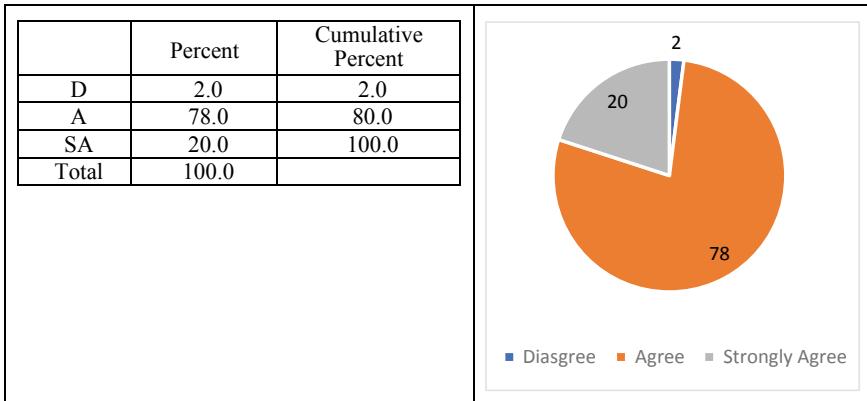
(vi). Do you feel the need of adequate external audits and security certifications to confirm cloud providers credibility? (Table 10)

Table 10 Questionnaire analysis**Part C**

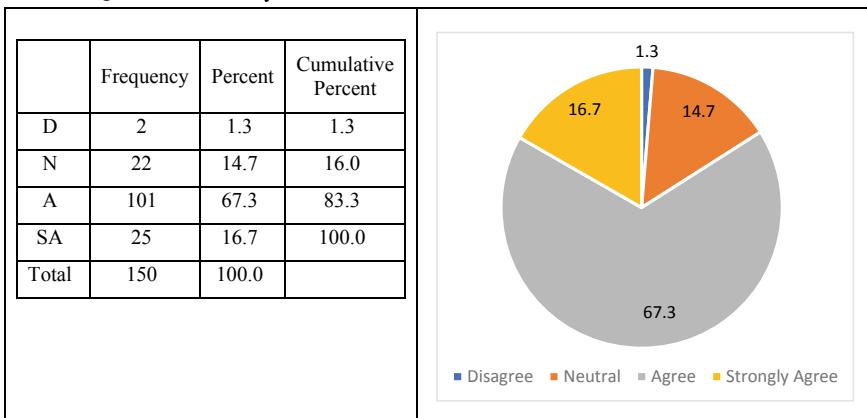
Q1 (i). Does the cloud provider provide an adequate mechanism to back your data? (Table 11)

Table 11 Questionnaire analysis

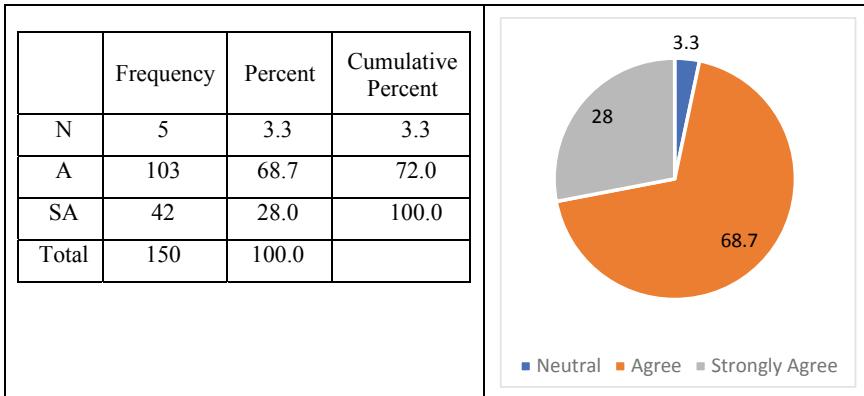
(ii). Is there a recovery mechanism for getting my backup? (Table 12)

Table 12 Questionnaire analysis

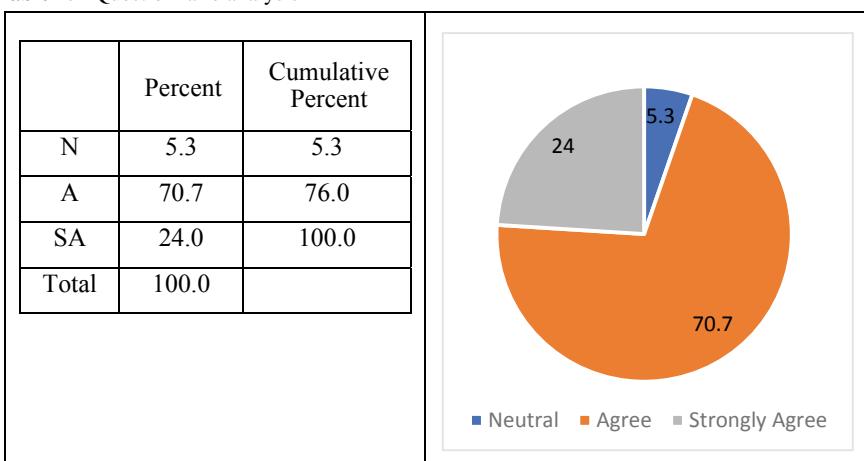
(iii). Does the Cloud provider hire people who are ethically trustworthy for managing and securing my data? (Table 13)

Table 13 Questionnaire analysis

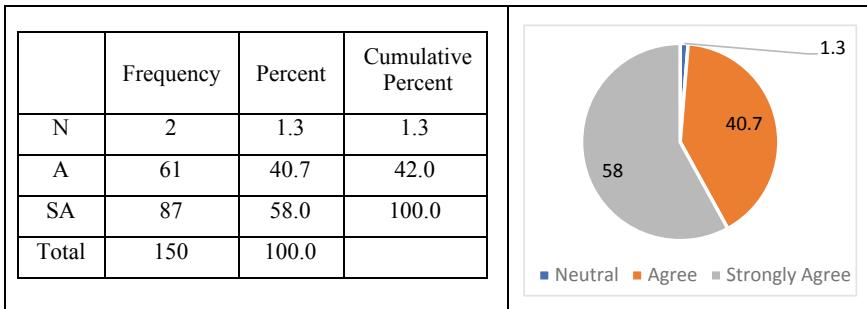
(iv). Does Cloud providers also have a moral duty to be honest with the customers regarding security policies and system architecture? (Table 14)

Table 14 Questionnaire analysis

(v). The word of mouth can create and influential impact on the consumers (Referees) (Table 15)

Table 15 Questionnaire analysis

(vi). The Cloud provider should be certified (Most cryptographic protocols for secure electronic transactions require the presence of trusted third party such as bank or certificate authority (CA)) (Table 16)

Table 16 Questionnaire analysis

3.2 Establishing Relationship Between Ethics and Trust in Cloud

User trust on the cloud on cloud services is governed by the experience of user in the field of work and average trust computed as a product of years of experience (N) and average of Cloud computing we considered two factors:

- Years of experience—N
- Average of Trust— A_T .

Trust is calculated as a product of Years of experience and Average of trust

$$N * A_T = \text{TRUST}.$$

From our questionnaire we have found out following data below in Table 17, it is not the complete data we are showing here few values from it (Table 18).

For estimating ethics, we consider:

- Designation of the respondent—P1
- Industry to which the respondent belongs—P2 (Table 19).

To calculate ethics based on

$$\text{ETHICS} = (P1 + P2/2) * A_E$$

where A_E is average of ethics computed from Table 17 ethic column values. Further to this correlation between Trust (X) and Ethics(Y) was calculated using Spearman's coefficient.

As per the Spearman Rank formula

$$(R) = 1 - \frac{6 \sum d^2}{n^3 - n}$$

$$(R) = 1 - (6 * 10)/(64 - 4)$$

$$(R) = +1$$

Table 17 Trust and ethics value computed based on questionnaire

Trust	Ethics	Trust	Ethics	Trust	Ethics	Trust	Ethics
3.667	4	4	4.43	4	4	3.5	4.29
4	4.71	4	4.43	4.667	4.43	4.333	4.29
4	4	4.333	4.29	4.333	5	4.167	3.86
3.333	4.29	3.5	4.29	3.5	4	3.333	4.14
4	4.71	4.167	4.29	4	4	4.333	4.43
4	4	3.833	4	3.833	4	3	4.29
4	4	3.167	3.43	3.5	4	4	4
3.167	3.43	3.333	4.29	4	4	3.667	4.14
3.333	4.29	3	3.86	4.333	4.43	4	4
3	3.86	4.5	5	4	4.14	4	4
4.5	5	4.5	5	3.667	4.14	4.667	4.43
4.5	5	4.5	4.86	4	4	4.167	4
4.5	4.86	4.667	4.57	4	4	3	4.43
4.667	4.57	5	5	3.667	4.29	3.667	4
5	5	4.167	3.86	4	4.29	4	4.43
4.333	5	3.333	4.14	4.333	4.14	4.333	4.29
4.167	4	4.333	4.43	4.333	3.86	3.5	4.29
3	4.43	3	4.29	3.5	3.57	3.667	4.29
3.667	4	4	4	3.5	3.57	4	4.29
4	4.43	3.667	4.14	3.333	4.14	4.333	4.14
4.333	4.29	4	4	4.333	4.43	4.333	3.86
3.5	4.29	4	4	3	4.29	3.5	3.57
4.333	4.29	4.667	4.43	4	4	3.5	3.57
4.167	3.86	4.5	5	3.667	4.14		
3.333	4.14	4.5	5	4	4		
4.333	4.43	4.5	4.86	4	4		
3	4.29	4.667	4.57	4.667	4.43		

Table 18 Normalized values of the experience factor

Years of experience	Value assigned
1–4 years	0.5
5–9 years	0.75
10 years above	1.0

Table 19 Normalized values for the designation of the respondent

Designation	Value assigned
Associate staff	0.25
Senior associate staff	0.5
Team leaders	0.75
Managers	1.0

A perfect positive correlation is $+1$ and a perfect negative correlation is -1 (Fig. 4).

It was observed that there lies a positive correlation between ethics and trust due to which the ethical behavior would lead to the development of trust on that organization thus boosting the commodity value of the organization.

To Find out a functional relationship between Trust and ethics we will use simple Linear Regression model. It is assumed that the two variables, trust (x) and ethics (y), are linearly related. Hence, we try to find a linear function that predicts the response value(y) as accurately as possible as a function of the feature or independent variable(x). Let us again use our dataset from Table 20 below which we have collected from different users from our questionnaire and we are a value of ethics as y for every trust value as x (Tables 21 and 22).

Now, our next task is to find a regression line which fits best in above values when we the scatter plot so that we can predict the response for any new feature values. The equation of regression line is represented as:

$$h(x_i) = \beta_0 + \beta_1 x_i$$

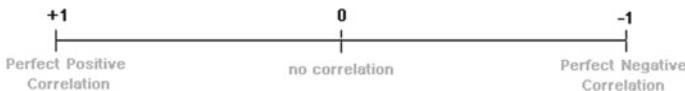


Fig. 4 Values of correlation

Table 20 Normalized values of the industry to which the respondent belongs

Industry	Value assigned
Software	1.00
Financial	0.75
Health care	0.60
Education	0.50
Gaming	0.30
Manufacturing	0.20
Others	0.10

Table 21 Ranking order of trust and ethics (sample values)

Trust (X)	Ethics (Y)	Rank (X)	Rank (Y)	D	d^2
1.80	3.00	3	2	1	1
2.00	3.53	2	1	1	1
3.00	2.50	1	3	-2	4
2.00	1.93	2	4	-2	4

Table 22 Values of Ethics and Trust calculated from the questionnaire

Trust	Ethics	Trust	Ethics	Trust	Ethics	Trust	Ethics
3.667	4	4	4.43	4	4	3.5	4.29
4	4.71	4	4.43	4.667	4.43	4.333	4.29
4	4	4.333	4.29	4.333	5	4.167	3.86
3.333	4.29	3.5	4.29	3.5	4	3.333	4.14
4	4.71	4.167	4.29	4	4	4.333	4.43
4	4	3.833	4	3.833	4	3	4.29
4	4	3.167	3.43	3.5	4	4	4
3.167	3.43	3.333	4.29	4	4	3.667	4.14
3.333	4.29	3	3.86	4.333	4.43	4	4
3	3.86	4.5	5	4	4.14	4	4
4.5	5	4.5	5	3.667	4.14	4.667	4.43
4.5	5	4.5	4.86	4	4	4.167	4
4.5	4.86	4.667	4.57	4	4	3	4.43
4.667	4.57	5	5	3.667	4.29	3.667	4
5	5	4.167	3.86	4	4.29	4	4.43
4.333	5	3.333	4.14	4.333	4.14	4.333	4.29
4.167	4	4.333	4.43	4.333	3.86	3.5	4.29
3	4.43	3	4.29	3.5	3.57	3.667	4.29
3.667	4	4	4	3.5	3.57	4	4.29
4	4.43	3.667	4.14	3.333	4.14	4.333	4.14
4.333	4.29	4	4	4.333	4.43	4.333	3.86
3.5	4.29	4	4	3	4.29	3.5	3.57
4.333	4.29	4.667	4.43	4	4	3.5	3.57
4.167	3.86	4.5	5	3.667	4.14		
3.333	4.14	4.5	5	4	4		
4.333	4.43	4.5	4.86	4	4		
3	4.29	4.667	4.57	4.667	4.43		
4	4	5	5	4.5	5		
3.667	4.14	4.333	5	4.5	5		

(continued)

Table 22 (continued)

Trust	Ethics	Trust	Ethics	Trust	Ethics	Trust	Ethics
4	4	4.167	4	4.5	4.86		
4	4	3	4.43	3.5	3.57		
4.667	4.43	3.667	4	3.667	4.29		
4.333	5	4	4.43	4	4.29		
3.5	4	4.333	4.29	4.333	4.14		

Here, $h(x_i)$ represents the predicted response value for i th observation. β_0 and β_1 are regression coefficients and represent y-intercept and slope of regression line, respectively. Next Step is to find out and estimate values of Coefficients B_0 and B_1 using least square technique:

$$\text{Consider } y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$\text{Now } y_i = h_{(x_i)} + \epsilon_i$$

$$\text{And } \epsilon_i = y_i - h_{(x_i)}$$

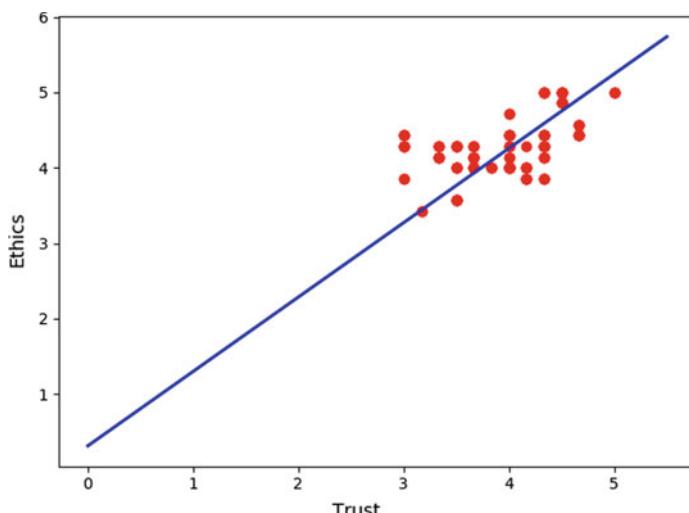
Here ϵ_i is residual error and our aim is to minimize the residual error.

$$\text{Square error or cost function is defined as } J(\beta_0, \beta_1) = \frac{1}{2N} \sum_{i=1}^n \epsilon_i^2$$

$$\text{where } \beta_1 = \frac{SS_{xy}}{SS_{xx}}$$

$$\text{where } SS_{xy} = \sum_{i=1}^n (x - \bar{x})(y - \bar{y}) \text{ and } SS_{xx} = \sum_{i=1}^n x^2 - n(\bar{x})^2$$

The relative values of Trust and Ethics was computed using Python and plotted as in Fig. 5 Output is coming out to be a straight line which indicates Trust and Ethics relationship is linear.

**Fig. 5** Regression between ethics and trust

3.3 Impact of Cloud Computing to Ecommerce Industry

The last section of the questionnaire above asked the respondents to rate the benefits of cloud computing a scale of 1–5 (5 strong positive impact and 1 means strong negative impact).

Descriptive statistics				
	Minimum	Maximum	Mean	Std. Deviation
1. Trust	3	5	4.49	0.576
2. Speed	3	5	4.26	0.607
3. Scalability	3	5	4.27	0.694
4. Cost	2	5	4.80	0.449
5. Interoperability	2	5	4.32	0.659
Valid N (list wise)				

From the above table it can be concluded that the most beneficial factor is cost with mean of 4.80 and low standard deviation of 0.449. Another important benefit is trust (mean = 4.49, SD = 0.576) followed by interoperability (mean = 4.32, SD = 0.659). High mean denotes that maximum respondents have chosen this factor as most beneficial and low standard deviation explains that there are low variations between the opinion of respondents. In other words, there is more consistency in the opinion of respondents.

4 Future Work

In any case, despite the limitations, as secured above, there are sure prospects which will aid the way toward structure trust in the cloud and which will shape the future extent of the work. The accompanying focuses are fundamental for prospects for creating trust in the cloud.

- Considering the above dataset, we can shape bunches for associations and discover the best one out of all.
- We can consider the effect of ethics and trust together on the web-based business associations utilizing the methods for clustering.

5 Conclusion

The paper examined the relationship between cloud computing and e-commerce. A model was developed to incorporate trust and ethics for an improved e-commerce environment, ETCM (Ethics trust cloud model). The new emerging technology of cloud computing is creating a new ecosystem service which will combine all the e-commerce services and facilitate the new service models. The survey conducted deduced the benefits of cloud computing on ecommerce and the findings of the survey and subsequent analysis lastly concluded that there lied a true positive correlation and regression between trust and ethics. A distinct method which provided a metric for ethical practices was proposed in the cloud environment. The work done thus proves a way to increase the quality and reliability of its work and services in cloud computing environment. The integration and collaboration of cloud computing open a new path for small and medium businesses and e-commerce can reach new heights.

References

1. G. Laatikainen, E. Luoma, Impact of cloud computing technologies on pricing models of software firms—insights from Finland, in *International Conference of Software Business, ICSOB 2014* (2014)
2. S.M. Habib, S. Hauke, S. Ries, M. Muhlhauser, Trust as a facilitator in cloud computing: survey. *J. Cloud Comput.* 1–19 (2012)
3. B. Hayes, Cloud Computing. *Commun. ACM* **51**, 9–11 (2008)
4. Haitao, Hangzhou to build China's first network service provider's cloud platform (2011)
5. <http://www.ufida.com.cn/>. Accessed June 2018
6. <http://www.ingdee.com/en>. Accessed June 2018
7. <http://www.eabax.com>. Accessed June 2018
8. S. Hupfer, Top ten ways cloud computing drives innovation (2014), <https://www.ibm.com/blogs/cloud-computing/2014/04/24/top-ten-ways-cloud-computing-drives-innovation/>
9. A. Greenberg, J. Hamilton, D. A. Maltz, P. Patel, The cost of a cloud: research problems in data center networks. *ACM SIGCOMM Comput. Commun. Rev.* 68–73 (2008)
10. S. Sajithabanan, E.G. Prakash Raj, Data storage security in cloud. *Int. J. Comput. Sci. Technol. (IJCST)* **2**(4), 1–5 (2011)
11. M. Armburst, A. Fox, R. Griffith, A.D. Joseph, R. Katz, G. Lee, Above the Clouds: A Berkeley View of Cloud Computing (2009), pp. 1–25
12. G. Dessler, B. Varkkey, *Human Resource Management*, 15th edn. (Pearson Education India, 2011)
13. M.A. Vouk, Cloud computing—issues, research and implementations. *J. Comput. Inf. Technol.* **16**(4), 235–246 (2004)
14. L. Wang, G. von Laszewski, A. Younge, X. He, M. Kunze, J. Tao, C. Fu, Cloud Computing: a perspective study **28**(2), 137–146 (2010)
15. W.-L. Chang, T.J. Allen, Amazon and Alibaba: Competition in Dynamic Environment (2016), <cnbc.com>. Accessed Sept 2017
16. N. Adyin, Cloud computing for E-commerce. *IOSR J. Mob. Comput. Appl.* **2**(1), 27–31 (2015)
17. H. Li, L. Tian, Y. Dai, H. Yang, Identity based Authentication for Cloud Computing (2009), pp. 157–166
18. P. Resnick, R. Zeckhauser, Trust among strangers in internet transactions: empirical analysis of eBay's reputation system. *Adv. Appl. Microecon. Res. Annu.* **11**, 127–157 (2002)

19. T. Grandison, M. Sloman, A survey of trust in Internet applications. *IEEE Commun. Surv. Tutor.* **3**, 2–16 (2000)
20. The benefits of cloud computing to Ecommerce industry **2**(1) (2015)
21. R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg, I. Brandic, *Future Generation Computer Systems*, vol. 25 (Elsevier, 2009), pp. 599–616
22. A. Sharma, H. Banati, A framework for implementing trust in cloud computing, in *Proceedings of the International Conference on Internet of things and Cloud Computing*, ACM, UK (2016)
23. A. Sharma, H. Banati, Ethical trust in cloud computing using fuzzy logic, in *Cloudcomp 2016* (Springer, China, 2017), pp. 44–55

Weighted Frequent Itemset Mining Using OWA on Uncertain Transactional Database



Samar Wazir, M. M. Sufyan Beg and Tanvir Ahmad

Abstract The technology of data mining has a broad scope in banking, manufacturing, medical science, and business decision-making. In these applications, the most commonly used term is Frequent Itemset Mining Algorithms which is one of the vital parts of Association Rule Mining. The evolutionary improvement in FIM is Weighted Frequent Itemset Mining, and these algorithms can be executed on Certain/Probabilistic Uncertain databases for calculating important frequent itemsets, having weight and expected support equal to or greater than user-specified minimum weight or minimum probability, respectively. The weight for an itemset is calculated by taking an average of weights of all items in the itemset. In this case, if the weights of one or two items are very high compared to others than the average can be decided by only higher weights and ignore low values. In this research paper, OWA operator is used in place of the traditional mean for calculating the weight of an itemset. A new algorithm is developed in this research and executed on example database. The results show that the generated itemsets are less in the count but hold more importance.

Keywords Certain/uncertain databases · Frequent items · Weighted mean · Expected support · OWA

1 Introduction

The dependency of a human being on technology is increasing day by day. Therefore, hardware/software used by different personal released a massive amount of data. This data is beneficial for future decision-making and refining this data according to the subject of use is known as data mining process. The data generated by the resources of the same category had some attributes in common and are linked to each other.

S. Wazir (✉) · T. Ahmad

Department of Computer Engineering, Jamia Millia Islamia, New Delhi 110025, Delhi, India
e-mail: samar.wazir786@gmail.com

M. M. Sufyan Beg

Department of Computer Engineering, Aligarh Muslim University, Aligarh, UP 202001, India

Table 1 Uncertain Database (UD)

TID	Items
1	A: 0.3, B: 0.6, D: 1.0
2	B: 0.7, F: 0.7
3	B: 0.9, C: 0.3, E: 1.0, F: 0.3
4	E: 1.0, F: 0.2
5	A: 0.5, C: 0.9, D: 0.7
6	A: 0.3, B: 0.6, C: 0.9, F: 0.6
7	A: 0.4, C: 0.9, D: 0.3, E: 1.0
8	C: 0.6, E: 0.4
9	A: 0.5, D: 0.7, F: 1.0
10	A: 0.7, B: 0.3, C: 0.8, E: 0.6

Finding out the data of the same class belongs to the Association Rule Mining (ARM) field of data mining. In a database, if some items occurred frequently imply that the items have some importance, therefore we need to extract such kind of items and this process known as Frequent Itemset Mining (FIM).

Data generated by different resources may be of various types, e.g. structured, unstructured, numeric, and alphanumeric, text, binary, etc. All kinds of data belong to either certain or uncertain category of databases and known as Certain Transactional Database (CD) or Uncertain Transactional Database (UD) here each row in the database can be said as a transaction. In CD, presence of an item in a row is definite on the other hand in UD the presence of an item in a row is probabilistic, e.g., in Table 1 there are 30% chances that item A is present in TID1 (Transaction ID). In UD each item is connected with its chances of existence, and this is called Existential Probability of item, e.g., in TID2, the Existential Probability of F is 0.7.

In transactional database, some items occurred frequently but are not important, and some occur rarely but are very important, e.g., if we take the example of Market Basket Data in which the profit of selling a toffee is 10 cents and selling a perfume is 100 dollars and on an average day sale of a toffee is 100 unit and perfume is 1 unit. In this case, perfume is more important than toffee. The importance of an item is decided by its weight and mining such items called Weighted Frequent Itemset Mining (WFIM). In Table 2, the items with their weights have demonstrated.

In WFIM, the weight of an itemset can be calculated by taking the average of weights of each item in the itemset, e.g., in Table 2, the weight of itemset ADF is 0.4 and CEF is 0.2. In itemset ADF the weight of items A and F is very small compare to D. In such a case, if we have some items in the database with very high weights (due to noise or error) then other items will be ignored while calculating the average. In

Table 2 Items weights

Items	A	B	C	D	E	F
Weight	0.1	0.8	0.2	1.0	0.3	0.1

this paper, the average is replaced by OWA, and a new algorithm has been proposed as Weighted Frequent Itemset Mining using OWA (WFIMOWA) which consider weights of all items in the itemset.

Paper is organised as follows: in Sect. 2 literature review related to research work has been presented. In Sect. 3 problem of calculating WFIM using average and OWA has been given. Section 4 consists of the proposed algorithm and its explanation. In Sect. 5, the problem of calculating WFIM using OWA has been demonstrated by using an example. In Sect. 6, the result of WFIM using average and OWA has been compared.

2 Related Works

The increasing demand for data analysis and data analytics leads toward the development of new strategies of FIM. The journey of FIM starts when Agrawal et al. [1] manifested an algorithm for FIM on CD in 1994 named as Apriori. This algorithm follows the sequential pattern of execution and consist of candidate generation-joining-pruning stages. When the size of itemsets increased, Apriori stuck in the problem of exponential candidate generation. Later, FP-Growth [2] and ECLAT [3] developed. The problem with CD is that imprecise, inaccurate and unstructured information cannot be stored in the CD. As a result, UD was introduced and the first algorithm for calculating frequent items on the uncertain database was developed by Chui [4, 5] as UApriori in 2007. UApriori also encountered two problems, first in case of lower minimum support the challenge of exponential candidate generation arises and second due to the multiplication of probabilities while calculating expected support, UApriori is unable to calculate long size frequent items. An alternative for UApriori was developed by Aggarwal as UHmine [6]. Later UFGrowth [7] and UECLAT [8] developed and introduced a new approach for mining FI by a tree-based and vertical mining approach, respectively.

At the end of twentieth-century mining of data become very popular and one of the most research-focused areas, consequently different algorithm has been developed by refining the older techniques. In fact, the only support of an itemset is not just the parameter for deciding that itemset is frequent or not but their confidence, correlation analysis, and weight also considered. The importance of an item is determined by the weight of that item in the database. On the basis of weight calculation, different algorithms have been developed for performing weighted frequent itemset mining, e.g. WAR [9], WARM [10], WFIM [11] and MWS [12] and for uncertain database U-WFI [13], HEWI-UApriori [14], WD-FIM [19].

3 Problem Statement and Solution

In this section, first, the calculation of OWA is explained then the problem of calculating WFIM using OWA is defined.

Consider the UD given in Table 1 and weights of items given in Table 2. Let $|D|$ is the size of UD in which $D = \{Tr_1, Tr_2, \dots, Tr_m\}$ consist of n no. of transactions and Tr_i is used to denote i th transaction. Let $X = \{x_1, x_2, \dots, x_m\}$ are the total number of items and $W(X) = \{w(x_1), w(x_2), \dots, w(x_m)\}$ are their corresponding weights in weight Table 2. $P_{x_{ij}}$ is used to denote the existential probability of j th item in the i th transaction of X itemset.

3.1 Calculation of Weight of an Itemset(X) Using Average

The weight of an itemset X can be calculated as

$$W(X)_{avg} = \frac{\sum_{i=1}^m w(x_i)}{m} \quad (1)$$

Here, m is the itemset size.

For example, in Table 2 $W(ABC)$ can be calculated as

$$W(ABC) = \frac{w(A) + w(B) + w(c)}{3} = 1.1/3 = 0.3667$$

3.2 Calculation of Weight of an Itemset X Using OWA [15–18]

The weight of itemset X using OWA can be calculated as

$$W(X)_{OWA} = \sum_{i=1}^m w_i y_i \quad (2)$$

Here, y_i is the i th largest value in x_1, x_2, \dots, x_m and w_i can be calculated as

$$w_i = Q\left(\frac{i}{m}\right) - Q\left(\frac{i-1}{m}\right), \quad i = 1, \dots, m \text{ with } Q(0) = 0$$

Here Q is called relative quantifier and can be calculated from

Table 3 Weights calculation for m = 2

i	0	1	2
i/m	0	0.5	1
Q	0	$\frac{(0.5-0.3)}{(0.8-0.3)}$	1
	0	0.4	1
W	NR	0.4	0.6

Table 4 Weights calculation for m = 3

i	0	1	2	3
i/m	0	0.33	0.66	1
Q	0	$\frac{0.03}{0.5}$	$\frac{0.36}{0.5}$	1
	0	0.06	0.72	1
W	NR	0.06	0.66	0.28

$$Q(r) = \begin{cases} 0 & \text{if } r < a \\ \frac{r-a}{b-a} & \text{if } a \leq r \leq b \\ 1 & \text{if } r > b \end{cases}$$

a and b are the relative linguistic fuzzy quantifier and in this paper $a = 0.3$ and $b = 0.8$ considered.

E.g. for $m = 2$ and $m = 3$, w_i can be calculated as in Tables 3 and 4.

Now, $W(ABC)_{OWA}$ can be calculated as

Here, $A = 0.1$, $B = 0.8$ and $C = 0.2$. Therefore, $y_i = 0.8, 0.2, 0.1$ and for $m = 3$.

$w_i = 0.06, 0.66, 0.28$ (as from Table 3).

So,

$$W(ABC)_{OWA} = \sum_{i=1}^m w_i y_i = 0.8 \times 0.06 + 0.2 \times 0.66 + 0.1 \times 0.28 = 0.208$$

3.3 Calculation of Expected Support [4, 5]

The expected support for an itemset X can be determined as

$$ES(X) = \sum_{i=1}^{|D|} \prod_{j=1}^{|m|} P_{x_{ij}}(X) \quad (3)$$

E.g., $ES(ABC)$ can be calculated as

$$\begin{aligned} ES(ABC) &= T1(P(ABC)) + T2(P(ABC)) + \cdots + T10(P(ABC)) \\ &= 0.162 + 0.168 = 0.33 \end{aligned}$$

3.4 Expected Weighted Support Using Average

Expected Weighted Support for an itemset X is the multiplication of itemset weight and its expected support. Therefore, from Eq. 1 and Eq. 3 $EWS(X)_{avg}$ can be calculated as

$$\begin{aligned} EWS(X)_{avg} &= W(X)_{avg} \times ES(X) = \frac{\sum_{i=1}^m w(x_i)}{m} \times ES(X) \\ EWS(ABC)_{avg} &= 0.3667 \times 0.33 = 0.121 \end{aligned}$$

3.5 Weighted Frequent Itemset Using Average

An itemset (X) is said to be weighted frequent if its expected weighted support using average is greater than or equal to user-specified minimum weighted support (mws). Let $mws = 0.1$. Then from Sect. 3.4

$$EWS(ABC)_{avg} = 0.121 \geq 0.1(mws)$$

So, ABC is a Weighted Frequent Itemset in this case where the average is used for weight calculation.

3.6 Expected Weighted Support Using OWA

Expected Weighted Support using OWA for an itemset X is the multiplication of itemset ordered weighted average ($W(X)_{OWA}$) and its expected support. Therefore, from Eq. 2 and Eq. 3 $EWS(X)_{OWA}$ can be calculated as

$$\begin{aligned} EWS(X)_{OWA} &= W(X)_{OWA} \times ES(X) \\ EWS(ABC)_{OWA} &= 0.208 \times 0.33 = 0.068 \end{aligned}$$

3.7 Weighted Frequent Itemset Using OWA

An itemset (X) is said to be weighted frequent using OWA if its expected weighted support using OWA is greater than or equal to user-specified minimum weighted support (mws). Let $mws = 0.1$. Then from Sect. 3.6

$$EWS(ABC)_{OWA} = 0.068 \leq 0.1(mws)$$

So, ABC is a not Weighted Frequent Itemset in this case where OWA is used for weight calculation.

4 WFIMOWA Algorithm

In the proposed algorithm, for new candidate generation, joining and pruning *UApriori*-like pattern is used.

WFIMOWA

1. Input (UD, WT, mws)
2. L1={large-1 itemset};//for all items in UD where $P_x \geq mws$
3. for (k=2; $L_{k-1} \neq \emptyset$; k++) do begin
4. $W(X)_{OWA} = OWA(k, WT)$
5. $C_k = \text{UApriori-gen } (L_{k-1})$ // New Candidates
6. $L_k = \{c \in C_k \mid c.\text{expectedSupport} \times W(X)_{OWA} >= mws\}$
7. End
8. Answer = $U_k L_k$

In the above algorithm, UD (uncertain database), WT (weight table), and mws (minimum weighted support) are given as input. In line-2, the existential probabilities of an item in all transactions are added and compared with mws . If it is greater than mws , then the item is added to the L_1 set. The process is repeated for all distinct items. In line-4 function, OWA takes k as criteria m and WT in input and return $W(X)_{OWA}$. In line-5 UApriori-gen function is used to join all size $k - 1$ items and perform pruning of candidates. Finally, size- k weighted frequent items using OWA is added to L_k set in line-6.

5 Example

Consider the database given in Table 1 and their weights provided in Table 2. Let mws given by the user is 0.1 then size-1, size-2, size-3 WFIM can be calculated as (Tables 5, 6 and 7).

Table 5 Size-1 WFIM

C1	ES	W	EWS	L1
A	2.7	0.1	0.27	A
B	3.1	0.8	2.48	B
C	4.4	0.2	0.88	C
D	2.7	1.0	2.7	D
E	4	0.3	1.2	E
F	2.8	0.1	0.28	F

Table 6 Size-2 WFIM using average and OWA

AVG					OWA				
C2	ES	W _{avg}	EWS _{avg}	L2	ES	W _{OWA}	EWS _{OWA}	L2	
AB	0.57	0.45	0.2565	AB	0.57	0.38	0.2166	AB	
AC	1.64	0.15	0.246	AC	1.64	0.14	0.2296	AC	
AD	1.12	0.55	0.616	AD	1.12	0.46	0.5152	AD	
AE	0.82	0.2	0.164	AE	0.82	0.18	0.1476	AE	
AF	0.68	0.1	0.068		0.68	0.1	0.068		
BC	1.05	0.5	0.525	BC	1.05	0.44	0.462	BC	
BD	0.6	0.9	0.54	BD	0.6	0.88	0.528	BD	
BE	1.08	0.55	0.594	BE	1.08	0.5	0.54	BE	
BF	1.12	0.45	0.504	BF	1.12	0.38	0.4256	BF	
CD	0.9	0.6	0.54	CD	0.9	0.52	0.468	CD	
CE	1.92	0.25	0.48	CE	1.92	0.24	0.4608	CE	
CF	0.63	0.15	0.0945		0.63	0.14	0.0882		
DE	0.3	0.65	0.195	DE	0.3	0.58	0.174	DE	
DF	0.7	0.55	0.385	DF	0.7	0.46	0.322	DF	
EF	0.5	0.2	0.1	EF	0.5	0.18	0.09		

Table 7 Size-3 WFIM using average and OWA

AVG		OWA								
C3	AP	ES	W _{avg}	EWS _{avg}	L ₃	C3	AP	ES	EWS _{owa}	L ₃
ABC	ABC	0.33	0.37	0.12	ABC	ABC	ABC	0.33	0.21	0.07
ABD	ABD	0.18	0.63	0.11	ABD	ABD	ABD	0.18	0.62	0.11
ABE	ABE	0.13	0.4	0.05	ABE	ABE	ABE	0.126	0.27	0.03
ACD	ACD	0.42	0.43	0.18	ACD	ACD	ACD	0.423	0.22	0.09
ACE	ACE	0.69	0.2	0.14	ACE	ACE	ACE	0.696	0.18	0.12
ADE	ADE	0.12	0.47	0.06	ADE	ADE	ADE	0.12	0.29	0.03
BCD	BCD	0	0.67	0	BCD	BCD	BCD	0	0.28	0
BCE	BCE	0.27	0.43	0.12	BCE	BCE	BCE	0.27	0.3	0.08
BCF	—	—	—	—	BCF	—	BCF	—	—	—
BDE	BDE	0	0.7	0	BDE	BDE	BDE	0	0.67	0
BDF	BDF	0.14	0.63	0.09	BDF	BDF	BDF	0.144	0.53	0.08
BEF	BEF	0.27	0.4	0.11	BEF	BEF	BEF	—	—	—
CDE	CDE	0.27	0.5	0.14	CDE	CDE	CDE	0.27	0.31	0.08
DEF	DEF	0	0.47	0	DEF	DEF	DEF	—	—	—

In the above table, the following terminologies are used
C: Candidate itemset, L: frequent itemset, AP: Candidate itemset after pruning

6 Result Analysis

In the analysis of FIM techniques, some parameters are considered, e.g., execution time taken by the algorithms and quality of generated frequent items. In the proposed algorithm, the value of $EWS(X)_{OWA}$ is always less than the value of $EWS(X)_{avg}$. So the number of candidates who qualify mws is always less in WFIMOWA. Therefore, the algorithm generates less number of candidates and faster in execution. Another parameter to be considered is, how important the generated frequent itemsets are. The importance of candidates can be checked by the following example.

Let's consider itemset ACD , which is frequent in $EWS(X)_{avg}$ but not in $EWS(X)_{OWA}$. In itemset ACD the weight of $D(1.0)$ is very high compared to $A(0.1)$ and $C(0.2)$, therefore, while calculating the average, A and C is almost ignored. D is high individually but when D is combined with A and C to form ACD itemset then the impact of A, C should also be considered. In other cases, ACE is frequent in both methods. Therefore, it can be said that the items generated by $EWS(X)_{avg}$ are important but the items generated by $EWS(X)_{OWA}$ are more important than average.

7 Conclusion and Future Directions

FIM is an essential technique to refine the business process, and various algorithms of FIM are available for producing results according to the user requirement. Sometimes number of frequent items are required irrespective of time and sometimes results are required in minimum time. In another case, sometimes the main focus is on the quality or importance of generated itemsets. The proposed algorithm generates more important itemsets in comparatively less time. The algorithm can be further extended by checking confidence and correlation between items to increase the quality of generated frequent items.

References

1. R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in *Proceedings of the 20th VLDB Conference*, Santiago, Chile (1994), pp. 487–499
2. J. Han, H. Pei, Y. Yin, Mining frequent patterns without candidate generation, in *Proceedings of Conference on the Management of Data* (SIGMOD'00, Dallas, TX) (ACM Press, New York, NY, USA, 2000)
3. M.J. Zaki, Scalable algorithms for association mining. *IEEE Trans. Knowl. Data Eng.* **12**(3), 372–390 (2000)
4. C.K. Chui, B. Kao, E. Hung, Mining frequent itemsets from uncertain data, in *11th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, PAKDD 2007*, Nanjing, China
5. C.K. Chui, B. Kao, A decremental approach for mining frequent itemsets from uncertain data, in *PAKDD* (2008), pp. 64–75

6. C.C. Aggarwal, Y. Li, J. Wang, J. Wang, Frequent pattern mining with uncertain data, in *Proceedings of ACM KDD* (2009), pp. 29–38
7. C.K.-S. Leung, M.A.F. Mateo, D.A. Brajczuk, A tree-based approach for frequent pattern mining from uncertain data, in *Proceedings of PAKDD* (2008), pp. 653–661
8. T. Calders, C. Garboni, B. Goethals, Efficient pattern mining of uncertain data with sampling, in *Proceedings of the PAKDD 2010, Part I* (Springer, 2010), pp. 480–487
9. W. Wang, J. Yang, P.S. Yu, Efficient mining of weighted association rules (war), in *Proceedings of 6th ACM SIGKDD International Conference on Knowledge Discovery Data Mining* (2000), pp. 270–274
10. F. Tao, F. Murtagh, M. Farid, Weighted association rule mining using weighted support and significance framework, in *Proceedings of 9th ACM SIGKDD International Conference on Knowledge Discovery Data Mining* (2003), pp. 661–666
11. U. Yun, J. Leggett, WFIM: weighted frequent itemset mining with a weight range and a minimum weight, in *Proceedings of SIAM International Conference on Data Mining* (2005), pp. 636–640
12. U. Yun, G. Lee, K.H. Ryu, Mining maximal frequent patterns by considering weight conditions over data streams. *Knowl.-Based Syst.* **55**(55), 49–65 (2014)
13. G. Lee, U. Yun, H. Ryang, An uncertainty-based approach: frequent itemset mining from uncertain data with different item importance. *Knowl. Based Syst.* **90**, 239–256 (2015)
14. A.C.-W. Lin, W. Gan, P. Fournier-Viger, T.-P. Hong, V.S. Tseng, Weighted frequent itemset mining over uncertain databases. *Appl. Intell.* **44**(1), 232–250 (2016)
15. R.R. Yager, Quantifiers in the formulation of multiple objective decision functions. *Inf. Sci.* **31**, 107–139 (1983)
16. R.R. Yager, On a general class of fuzzy connectives. *Fuzzy Sets Syst.* **4** (1980)
17. R.R. Yager, On ordered weighted averaging aggregation operators in multicriteria decision making. *IEEE Trans. Syst. Man Cybern.* **18** (1988)
18. S. Wazir, M.M. Sufyan Beg, T. Ahmad, Frequent itemset mining on uncertain database using OWA operator, in *Proceedings of 2nd International Conference on Communication, Computing and Networking*, ed. by C. Krishna, M. Dutta, R. Kumar. Lecture Notes in Networks and Systems, vol. 46 (Springer, Singapore, 2019)
19. X. Zhao, X. Zhang, P. Wang, S. Chen, Z. Sun, A weighted frequent itemset mining algorithm for intelligent decision in smart systems. *IEEE Access* **6**, 29271–29282 (2018)

Design of Customer Information Management System



Rohini Narayan and Gitanjali Mehta

Abstract The Customer Information Management (CIM) needs to handle, maintain and store the customer data. To facilitate this objective, the need for an application that takes the customer and the item details from the external source files and store the same in the Data Warehouse (DW) is required. It is an ETL system developed to create the mapping to load the target from the source files after applying criteria as required by the administrator and will also provide the overall features and functionalities required to store and maintain customer data. The data stored in the DW is transformed (cleaned and integrated), thus it is credible and can be used for business insights. This serves as an important component for BI (Business Intelligence) which will help in transforming raw/operational data into some meaningful information. The Fact_Inventory which is developed provides us the historical data in a summarized form which can be used for managerial, strategic and analytical decisions by the Analysts team and the end users. It acts as a vast storehouse for the already operated data and can be referred to as an 'Informational System'.

Keywords Information management · Data warehouse · Business intelligence

1 Introduction

1.1 Overall Description

The aim of the work is to create a Data Warehouse for Customer Information Management using ETL (Extract, Transform, and Load) where we can store all the major customer details regarding Customer ID, Average cost, Total count of sales, etc. on monthly basis. A data warehouse is constructed by integrating data from multiple heterogeneous sources (Flat files, dBs, mainframes, etc.) that support analytical reporting, structured and/or ad hoc queries, and decision-making. It is not loaded

R. Narayan · G. Mehta (✉)

School of Electrical, Electronics and Communication Engineering, Galgotias University, Greater Noida, India

e-mail: gitanjali.iitr@gmail.com

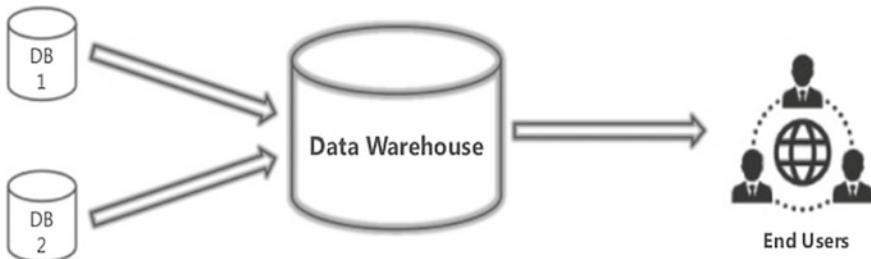


Fig. 1 DW system

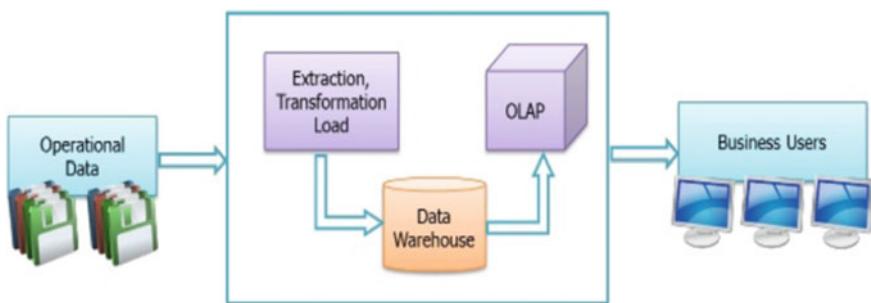


Fig. 2 Approach of DW

every time new data is added to the database and is accessed by the end users only when information is needed (Fig. 1) [1–3].

It gives us the summarized data for our customers and corresponding item sales which can be helpful for sample survey. It is a relational database that is designed for query and analysis. It usually contains historical data derived from transaction data and other sources. Before being stored in the data warehouse, the data undergoes transformation and cleansing to be used in the standard format. While operational data is organized by specific processes or tasks and is maintained by separate systems, warehoused data is organized by subject area and is populated from many operational systems (Fig. 2) [4, 5].

1.2 Purpose

The Data Warehouse constructed will help us to meet strategic requirements for customer information regarding customer behavior towards sales, item preferences, etc. It is a blend of technologies and components which allows the strategic use of data. Moreover, the nonvolatile nature of DW enables the previous data to be retained even when new data is added to it; thus allowing us to study patterns over historic

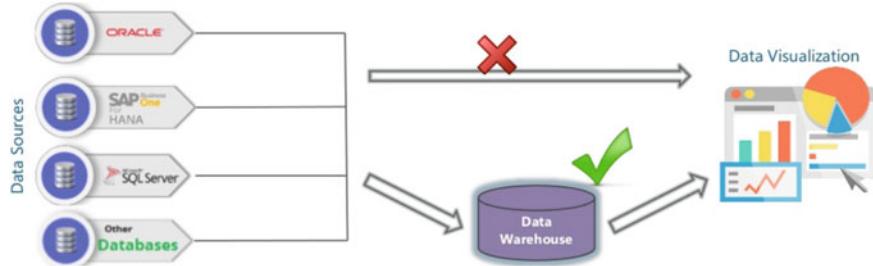


Fig. 3 Data visualization through DW

data and make the necessary predictions [6–8]. The strategic data contained in CIM Fact_Inventory helps the business with the following needs:

- Understand Customer Behaviour
- Analyze Trends and Relationships
- Analyze Problems Of Sales
- Discover Business Opportunities For Individual Items
- Plan for the Future

The data collected from various sources and stored in various databases cannot be visualized directly. The data first needs to be integrated and then processed before visualization takes place. The CIM serves this purpose by providing us with the Data Warehouse System which provides an excellent approach for transforming the vast amounts of data into reliable information that can support the decision-making process (Fig. 3) [9].

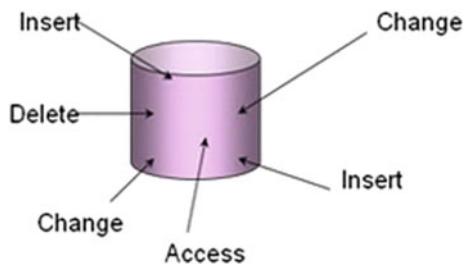
1.3 Motivations and Scope

The motivation behind this work is the concept of Data Warehousing, where we collect the data from dissimilar sources and then clear it; organize it to be supported by BI. It allows business users to quickly access critical data from varied sources all in one place.

Normally data collected from various sources and stored in various databases cannot be visualized directly. This calls for the need to create a Fact_Inventory which is the Data Warehouse for our Customer Information Management (CIM). It contains the summarized data for our customers and items on an aggregate basis.

The scope of our ETL process is limited to data load and data access. Data is in the read-only format and is periodically refreshed. It is stored as snapshots, each representing a period of time. This helps to analyze historical data and understand what and when happened. Thus, our focus is not on the ongoing operations, rather it is on the modeling and analysis of data for decision-making (Fig. 4).

Fig. 4 Transactional data storage



There is no such issue in the existing system. It is just having a limited scope where tasks regarding managerial decisions, strategic analysis, etc. cannot be done. Also data collected from various sources like flat files, servers, dBs, mainframes, etc. don't prove to be beneficial for the visualization purpose. This calls for the need of integration where the establishment of a common unit of measure for all similar data from the dissimilar database takes place after which the necessary data load can be done.

This paper presents the ETL process for building a Data Warehouse that keeps the summarized data of customer sales. The emphasis is on the development of a centralized data access (Fact_Inventory) for generating various reports to forecast the items' sales corresponding to individual customers based on their monthly purchase count, average cost, etc.

2 Proposed Model

2.1 System Environment

It gives us the layout of DW under construction. We get an overall idea about the flow of data from the sources to the targets. The various tables involved in the process are discussed here (Table 1).

Table 1 CIM system environment

Table name	Load type	Table description
Customer	Type 2 dimension	Contains all the information of the customer
Item	Type 2 dimension	Contains information about the item
Date	Reference table	Contains information regarding the date
Customer_item	Bridge table	Contains the data which is related to the customer and the item
Fact_inventory	Fact table	Contains the final transformed data

- The dimension tables involve the concept of SCD (Slowly Changing Dimension). In CIM we have implemented the SCD-Type 2 where all history of dimension changes is kept in the database. We capture attribute change by adding a new row with a new surrogate key to the dimension table. Both the prior and new rows contain as attributes the natural key (or other durable identifiers). Here the difference in the End Date of the updated record helps us in identifying the necessary changes and using the latest information for analytical purpose.
- Date Table is used as the Reference Table where the date_id can be used for monthly data storage in the Fact_Inventory.
- The Customer_Item table acts as the Bridge Table which acts as an intermediate between the Dimension and the Fact. It contains the keys of the corresponding dimension tables and is used to resolve many-to-many relationships between a fact and a dimension.
- The concept of surrogate keys has also been utilized for the sequential assignment of records during the dimension load. This proves to be extremely useful for analytical purpose. Since surrogate keys are system-generated, it is impossible for the system to create and store a duplicate value. It improves the performance of the Fact table as most of its attribute types are foreign keys.
- The CIM works on the Star Schema where Fact Table (holding the foreign keys column that allows joins with dimension tables, and the measures columns containing the data that is being analyzed) sits at the center with the dimension tables at its corners.

2.2 *System Architecture*

The Customer Information Management consists of developing an ETL process with the Fact table as the central repository of data where the aggregate information for customer purchases can be utilized for decision-making purpose. The various phases through which data passes before finally being stored in the Fact table are Staging, Dimension/Bridge, etc. This is depicted through the architecture diagrams as under:

High-Level Architecture Diagram (Fig. 5)

Here the flow of data is shown on an overall level from sources (Dim_Cust, Dim_Item, and Cust_Item) to the target, i.e.; Data Warehouse via the Staging Area. In the first half, data is truncated and loaded into the Staging Area while in the second half, it undergoes necessary transformation before being loaded into the target. During the ‘Truncate and Load’ the unwanted data will go to the Bad Files and only the useful data will be passed on to the Staging Area. In case of ‘Transform and Load’, the staged data passes through router, joiner, expressions, etc., before being loaded in the dimension tables. In the end, the transformed data passes through the aggregator function before being finally stored into the Data Warehouse where it is utilized for generating the reports and making decisions for analytical purpose.

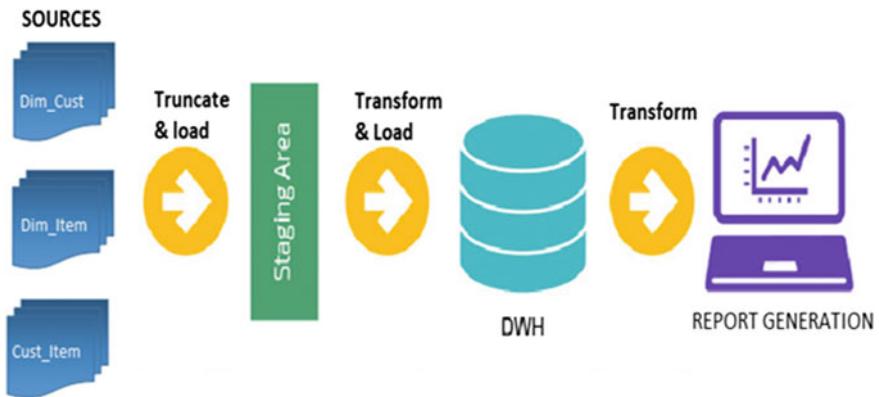


Fig. 5 High-level architecture diagram

Low-Level Architecture Diagram (Fig. 6)

The flow of data is shown schematically from the Source Files (Customer and Item) and Cust_Item (Bridge Table) which truncate and load the data into the Staging Area. This acts as the intermediate storage area for data which sits between the Source and the Target. At this level, we perform cleansing, scrubbing, etc. The data then undergoes necessary transformations after which it is loaded into the respective

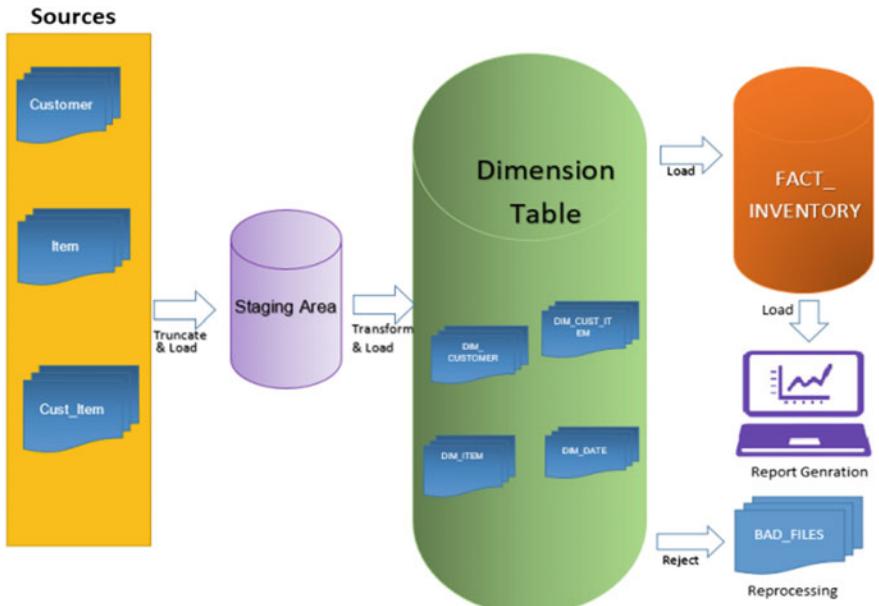


Fig. 6 Low-level architecture diagram

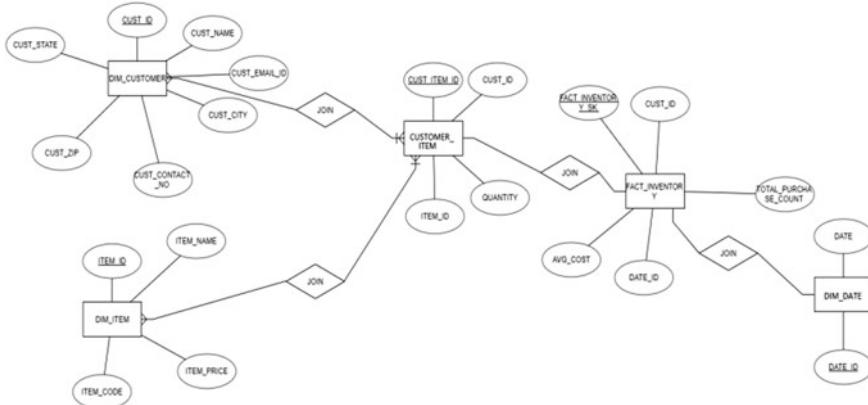


Fig. 7 E-R diagram

dimension tables. From here the data either passes to the Fact_Inventory after which the report is generated by the analyst team or is rejected into the Bad Files for reprocessing.

Entity-Relationship Diagram (Fig. 7)

The E-R Diagram shows 5 different entities, i.e.; Dim_Customer, Dim_Item, Customer_Item, Fact_Inventory, and Dim_Date. The entities are related through the Join Relationship where the Bridge Table loads the data from the respective dimension tables based on the values of the keys. The data is then loaded into the Fact_Inventory which is also mapped with the Dim_Date to generate reports of the customers based on the monthly purchase which can be referred later for decision-making.

Use-Case Diagram

See Fig. 8.

2.3 Methodology, Tools, and Techniques

Here, we apply the Agile Software Development Life Cycle (SDLC) Model to implement our project. The development of CIM Fact_Inventory is facilitated with the help of Oracle dB for queries processing and Informatica Power Center Tool for the ETL process (Fig. 9).

Oracle 11g Express Edition

It is a multi-model database management system produced and marketed by Oracle Corporation. It is a database commonly used for running online transaction processing (OLTP), data warehousing (DW) and mixed (OLTP & DW) database workloads (Fig. 10).

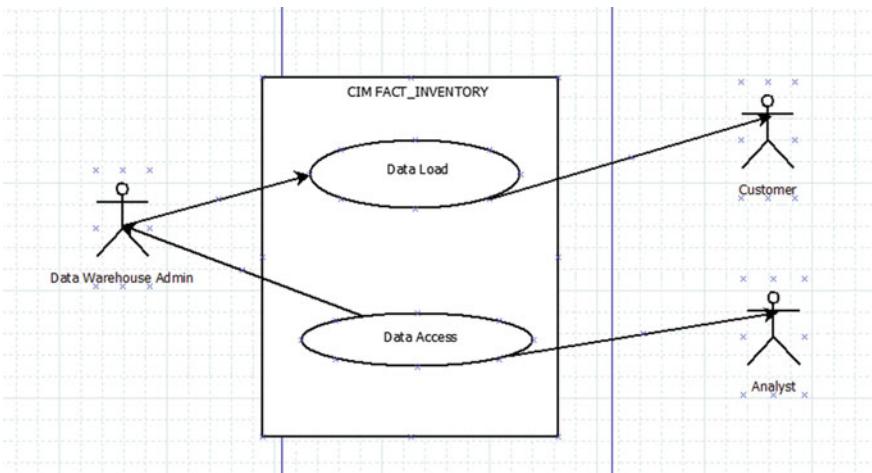
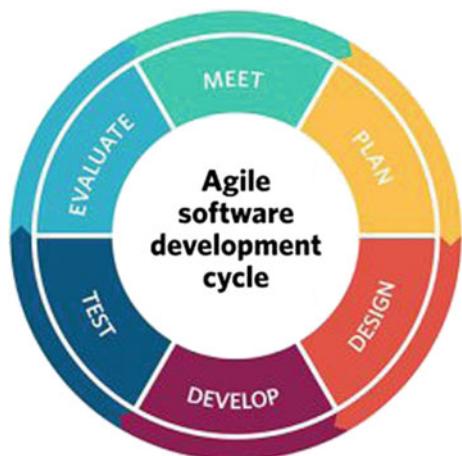


Fig. 8 Use-case diagram

Fig. 9 Agile software development cycle



Informatica Power center client 10.1

It is a graphical user Interface used to build and manage PowerCenter objects like source, target, Mapplets, Mapping, and transformations. It has a set of tools that are used to design ETL applications called “Mapping” (Fig. 11).

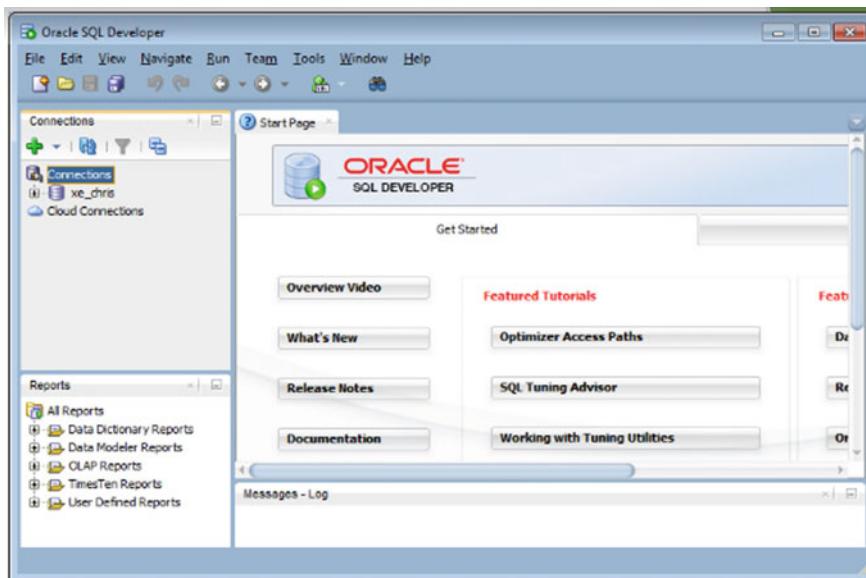


Fig. 10 Layout of Oracle SQL developer window

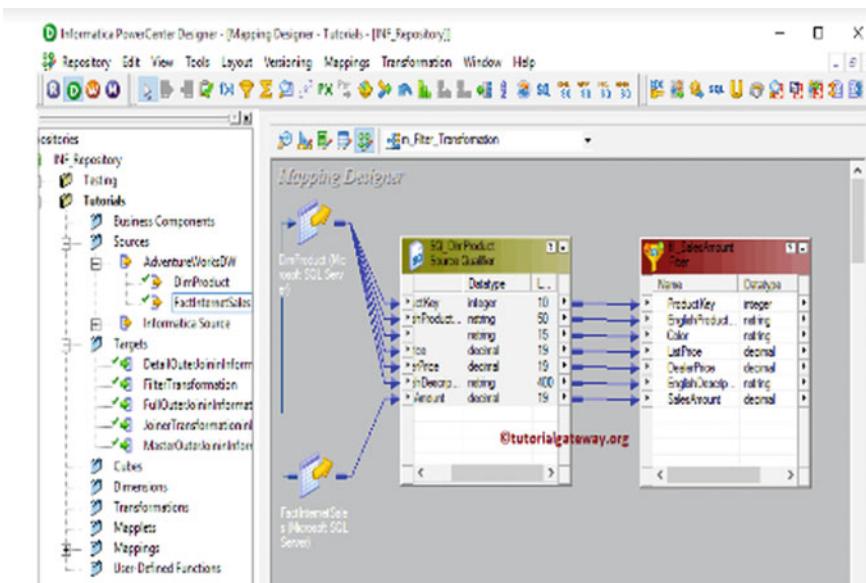


Fig. 11 Layout of Informatica power center designer

Functional Requirements

Customer information management	
Functional requirements-1 (Stage Load)	<ul style="list-style-type: none"> Capture data from the customer and item source files Load them into the landing zone, i.e., staging area Apply cleansing, scrubbing, etc. to remove the duplicated data, non-formatted data, etc.
Functional requirements-2 (Dimension Load)	<ul style="list-style-type: none"> Load the staged data into the Dim_Customer and Dim_Item All the necessary transformations take place at this level
Functional requirements-3 (Fact Load)	<ul style="list-style-type: none"> Populate the Fact_Inventory with the summarized data of the customers and their corresponding sales on a monthly basis Acts as the DW for decision-making purpose

3 Implementation

File To Stage Load

- The Extract part of ETL process takes place here where we collect data from various sources and load them in the Staging area.
- Src_Customer and Src_Item are the flat files from which the necessary data of the customers and items are collected, respectively, for our CIM.
- The Data Staging Area sits between the data sources and data targets.
- It prepares the crucial data before loading it into the dimensions.
- Tasks like cleansing, scrubbing, truncate, etc., to remove the unwanted data take place at this phase.
- Thus, data is converted to the standardized format in the Stg_Customer and Stg_Item.
- This proves to be significant as it is used to quickly extract data from the sources, minimizing the impact of the sources (Figs. 12, 13 and Table 2).

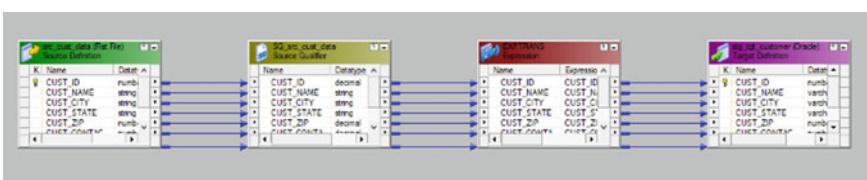


Fig. 12 Mapping from Src_Customer to Stg_Customer (m_load_stg_customer)



Fig. 13 Mapping from Src_Item to Stg_Item (m_load_stg_item)

Stage to Dimension Load

- This plays a vital role in the ETL process. The second and most crucial phase of data transformation takes place here.
- Data Type conversion, Joiners, Expressions, etc. are performed to avoid any data type mismatch.
- Business Rules are applied during the Dimension load.
- Also parent-child relationships are maintained.
- Special care is taken of date format which needs to be changed to a standardized format.
- The concept of SCD (Slowly Changing Dimension) is applied here on the dimension tables. In CIM we have implemented the SCD-Type 2 where all history of dimension changes is kept in the database. We capture attribute change by adding a new row with a new surrogate key to the dimension table. Both the prior and new rows contain as attributes the natural key. Here the difference in the End Date of the updated record helps us in identifying the necessary changes and using the latest information for analytical purpose.
- Also, we make use of Surrogate Keys which is generated by making use of Sequence Generator to avoid any confusion and duplicate values.
- CIM works on ‘Star Schema’ where the fact table sits in the center with the dimensions on its corners (Figs. 14, 15, 16 and Table 3).

Dimension to Fact Load

- The transformed data collected in the Dim_Customer and Dim_Item are then sent to the Fact_Inventory for the final load.
- The integrated data is loaded into the presentation area of the data warehouse.
- Loading is basically the process of loading the data in the data warehouse so that it can be used for the analytical purpose. Indexing should be there in the DW before the arrival of data for better query performance.
- The Fact_Inventory in the CIM consists of the aggregate data of the customers which give us the summarized details of the sales on a monthly basis.
- A fact table stores quantitative information for analysis and is often denormalized.
- The Customer_Item table acts as the Bridge Table which acts as an intermediate between the Dimension and the Fact. It contains the keys of the corresponding dimension tables and is used to resolve many-to-many relationships between a fact and a dimension.

Table 2 Source to stage mapping table

SRC_NAME	SRC_NAME	SRC_NAME	SRC_NAME	SRC_NAME	TGT_COL
SRC_CUSTOMER	CUSTOMER_ID	LTRIM(RTRIM(customer_name))	STG_CUSTOMER	STG_LOAD	CUSTOMER_ID
	CUSTOMER_NAME	LTRIM(RTRIM(customer_city))			CUSTOMER_NAME
	CUSTOMER_CITY	LTRIM(RTRIM(customer_state))			CUSTOMER_CITY
	CUSTOMER_STATE	LTRIM(RTRIM(customer_zip))			CUSTOMER_STATE
	CUSTOMER_ZIP				CUSTOMER_ZIP
	CUSTOMER_CONTACT_NO				CUSTOMER_CONTACT_NO
	CUSTOMER_EMAIL_ID	LTRIM(RTRIM(customer_email_id))			CUSTOMER_EMAIL_ID
STG_ITEM	ITEM_ID	STG_ITEM	STG_LOAD	ITEM_ID	ITEM_ID
	ITEM_NAME	LTRIM(RTRIM(item_name))			ITEM_NAME
	ITEM_PRICE				ITEM_PRICE
	ITEM_CODE				ITEM_CODE

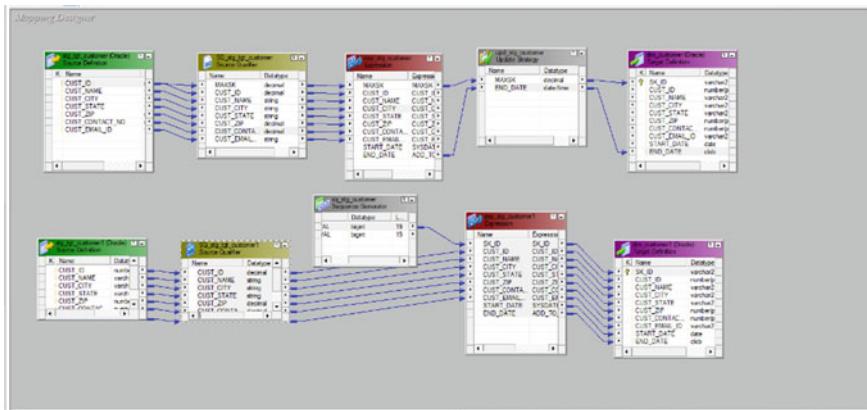


Fig. 14 Mapping from Stg_Customer to Dim_Customer (m_load_dim_customer)

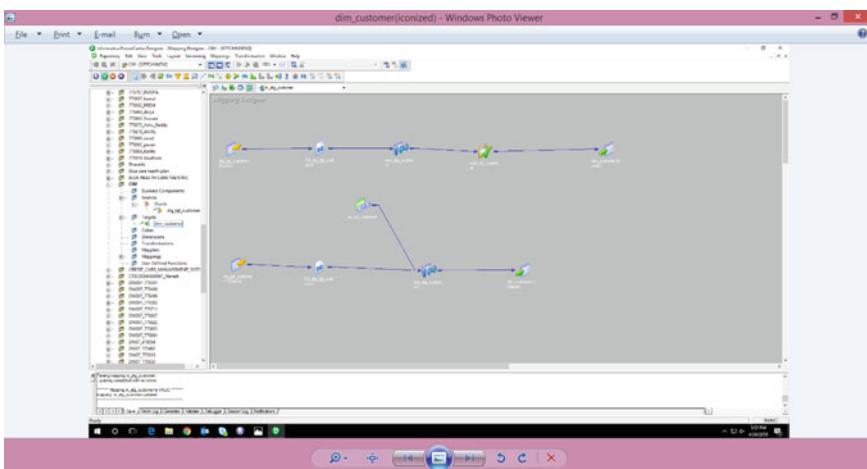


Fig. 15 Iconized Mapping from Stg_Customer to Dim_Customer (m_load_dim_customer)

- System-generated surrogate keys are used which avoid duplicate values and improve the performance of Fact table as most of its attribute types are foreign keys.
- The data stored in the DW can act as a source for decision-making (Figs. 17, 18 and Table 4).

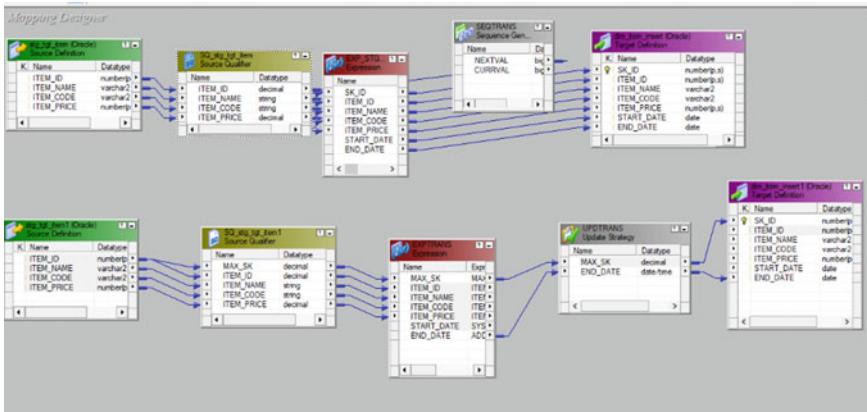


Fig. 16 Mapping from Stg_Item to Dim_Item (m_load_dim_item)

4 Results and Discussions

Sample Result (Figs. 19 and 20)

- In CIM, we are mainly concerned with the creation of a Fact_Inventory (DW) which contains the summarized data of the Customers and their corresponding sales/purchases of items.
- SCD-Type 2 becomes essential as we have to retain the history of all records so that we can easily make out the time period of the changes. Also, crucial data are maintained in the Customer and the Item dimensions which should not be lost.
- Upon completing with the creation of Fact_Inventory the stored data need not be changed on a frequent basis. It is added regularly, but loaded data are rarely changed directly.

Test Cases

We apply Unit Testing on our Customer Information Management (CIM). Testing is done on every level to ensure that the correct data in the standard format passes to the next level.

Unit Testing For Source to Stage Load

See Table 5.

Unit Testing For Stage to Dimension Load

See Table 6.

Unit Testing for Dimension to Fact Load

See Table 7.

Table 3 Stage to dimension mapping table

SRC_NAME	SRC_COL	TRANSFORMATION	LOAD TYPE	TGT_TB	TGT_COL
STG_CUSTOMER	CUSTOMER_ID	TO_INTEGER(customer_id)	Type 2	CIM_DIM_CUSTOMER	CUSTOMER_ID
	CUSTOMER_NAME			CUSTOMER_NAME	
	CUSTOMER_CITY			CUSTOMER_CITY	
	CUSTOMER_STATE			CUSTOMER_STATE	
	CUSTOMER_ZIP			CUSTOMER_ZIP	
	CUSTOMER_CONTACT_NO			CUSTOMER_CONTACT_NO	
	CUSTOMER_EMAIL_ID			CUSTOMER_EMAIL_ID	
	SYSDATE			START_DATE	
	NULL			END_DATE	
	ITEM_ID	TO_INTEGER(item_id)	Type 2	CIM_DIM_ITEM	ITEM_ID
STG_ITEM	ITEM_NAME			ITEM_NAME	
	ITEM_PRICE			ITEM_PRICE	
	ITEM_CODE			ITEM_CODE	

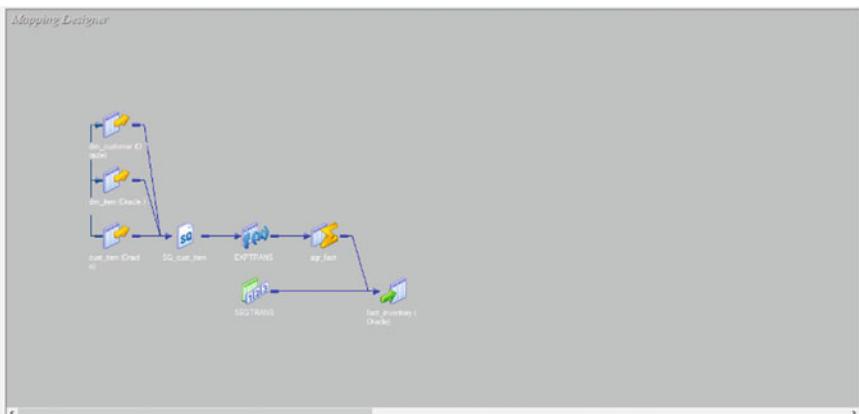


Fig. 17 Iconized mapping from dimensions to fact (m_load_fact_inventory)

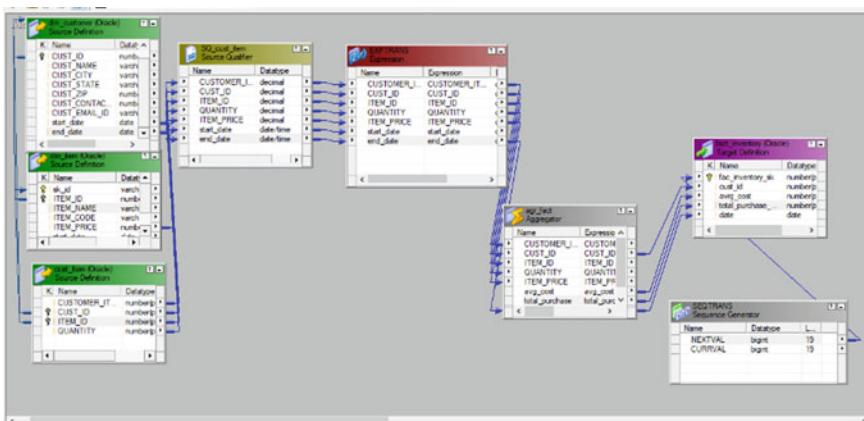


Fig. 18 Mapping from dimensions to fact (m_load_fact_inventory)

5 Conclusion

Customer Information Management (CIM) provides us with the ETL process to construct a Data Warehouse, a central repository for storing the summarized data of customers' sales, which can be utilized by the Data Analysts or the end users for making strategic decisions. The conventional dBs are merely used only for operating on the current data. However, by developing a DW we can store all the historic data and use them in forecasting the future business plans by studying the various patterns in customer behavior. The CIM Fact_Inventory serves this purpose by storing the aggregate values. It holds the foreign keys column that allows joins with dimension tables and the measures columns containing the data that is being analyzed. Thus,

Table 4 Dimension to fact load

SRC_NAME	SRC_COL	TRANSFORMATION	TGT_TB	TGT_COL
DIM_CUSTOMER	CUSTOMER_ID	TO_INTEGER(customer_id)	FACT_INVENTORY	CUSTOMER_ID
	SYSDATE		START_DATE	
	NULL		END_DATE	
DIM_ITEM	ITEM_PRICE			
DIM_CUST_ITEM	QUANTITY	QUERY	AVG_COST	
		QUERY	TOTAL_PURCHASE_COUNT	
		PRIMARY KEY	FACT_INVENTORY_SK	

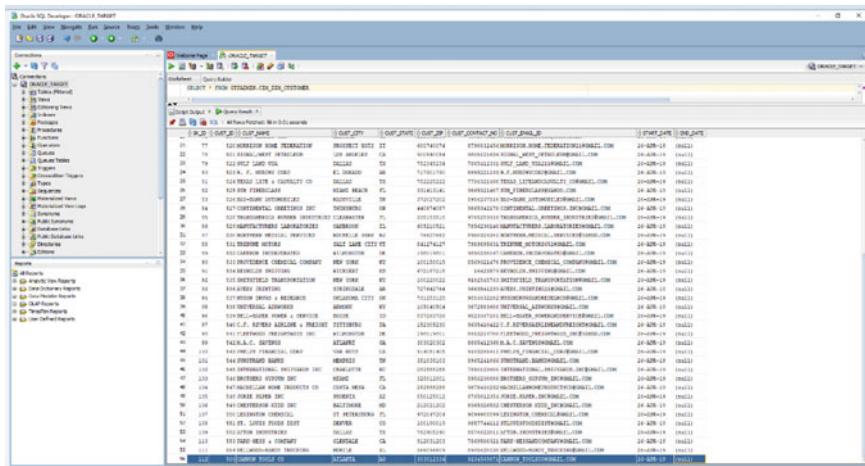


Fig. 19 Sample output after SCD2

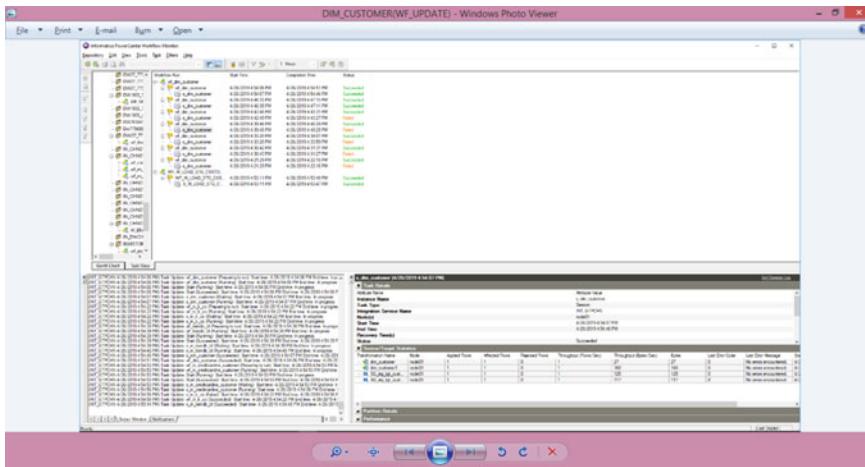


Fig. 20 Workflow and sessions log window for Dim_Customer

our purpose of developing an information system in the form of Data Warehouse is fulfilled.

Table 5 Unit testing table for source to stage load

Test case ID	Test case	Test case type	Steps to execute the test case	Expected result	Actual result	PASS/FAIL
1.	Count of Source file and Stage table	Positive	No. of rows in the Source file should match the data truncated and loaded in the Stage Tables	No. of rows in the Source Customer/Item File = No. of rows loaded in the Customer/Item Stage Table	No. of rows in the Source Customer/Item File = No. of rows loaded in the Customer/Item Stage Table	PASS
2.	Verify random 5 records data both in source file and stage	Positive	The data in the Stage table should be matched with the corresponding Cust_Id/Item_Id from the Source File	Data in the Source File should match Data in the Stage Table	Data in the Source File matches Data in the Stage Table	PASS

Table 6 Unit testing table for the stage to dimension load

Test case ID	Test case	Test case type	Steps to execute the test case	Expected result	Actual result	PASS/FAIL
1.	Old records being loaded into dimension as insert and update	Positive	In case of the single row being updated the no. of rows in the dimension, table should be one more than the ones in the Stage table	No. of rows in the Stage Customer/Item Table = $1 + \text{No. of rows loaded in the Customer/Item Dimension Table}$	No. of rows in the Stage Customer Table = $1 + \text{No. of rows loaded in the Customer Dimension Table}$	PASS
2.	New records being loaded into the dimension as insert	Positive	In case of new rows, the no. of rows in the dimension table should be same as the ones in the Stage table	No. of rows in the Stage Customer/Item Table = $\text{No. of rows loaded in the Customer/Item Dimension Table}$	No. of rows in the Stage Customer/Item Table = $\text{No. of rows loaded in the Customer/Item Dimension Table}$	PASS

Table 7 Unit Testing table for dimension to fact load

Test case ID	Test case	Test case type	Steps to execute the test case	Expected result	Actual result	PASS/FAIL
1.	Rows being inserted into the Fact on a monthly basis	Positive	The data from the dimension should be loaded into the Fact_Inventory aggregated on a monthly basis	No. of rows in the Dimension Customer and Item Tables should match with the corresponding monthly basis data loaded in the Fact_Inventory	No. of rows in the Dimension Customer and Item Tables matches with the corresponding monthly basis data loaded in the Fact_Inventory	PASS

References

1. N.L.B. Putri, Design of information systems customer relationship management to improve services sales approach system development life cycle (SDLC), **6**(1), 466–472 (2017)
2. H.S. Soliman, Customer relationship management and its relationship to the marketing performance. *Int. J. Bus. Soc. Sci.* **2**(10), 166–182 (2011)
3. M. Anshari, N. Almunawar, M. Arif, S. Lim, A. Al-Mudimigh, Customer relationship management and big data enabled: personalization & customization services. *Appl. Comput. Inform.* **15**(2), 94–101 (2019)
4. E. Prayitno, N.A. Astuty, Positive impact of customer relationship management (CRM) implementation to improving the services of animal polyclinics customers, in *2017 International Conference on Sustainable Information Engineering and Technology (SIET)*, Malang, 2017, pp. 246–250
5. W.K.R. Perera, K.A. Dilini, T. Kulawansa, A review of big data analytics for customer relationship management, in *International Conference on Information Technology Research*, 2018, pp. 1–6
6. I.J. Chen, K. Popovich, Understanding customer relationship management (CRM) people process and technology. *Bus. Process Manag. J.* **9**(5), 17, 672–688 (2003)
7. G. Mehta, G. Mittra, V.K. Yadav, Application of IoT to optimize data center operations, in *International Conference on Computing, Power and Communication Technologies*, 2018, pp. 738–742
8. A. Mishra, D. Mishra, Customer relationship management: implementation process perspective. *Acta Polytech. Hung.* **6**(4), 83–99 (2009)
9. M. Srivastava, Customer relationship management: a technology driven tool. *SIBM* **2**, 14–25 (2012)

Modeling Machine Learning Agent for Interaction Conversational System Using Max Entropy Approach in Natural Language Processing



Anil Kumar Negi and Syed Imtiyaz Hassan

Abstract There are several service-oriented models, where services are deployed for the user. The user chooses any one of them. During the operational life cycle of these services, there are several issues that occur. User wants an interface for complaint. This paper uses the sentence boundary detection, NER, document categorization, and sentiment extraction methodologies. The natural language processing generates the training model which is a statistical representation of current system knowledge. When a user enters the input then training model extracts the value of a different parameter, these parameters used by call center model for better understanding of user input. The machine learning model used to generate a logical response. When the time exceeds the size of sample, data should be increased and the model understanding also more and more accurate. The accuracy of the model depends on the size of the sample training data. When training data size increases the statistical model for the call center is updated. When the user interacts with the call center agent then call center agent to extract the value of all parameters based on the current statistical model which is based on the sample training data. The paper uses different parameter such as NER, Document category, and sentiment for making a better user interaction. The probability of correct response is increase n time if n parameters are used for response generation. Call center module to take help from sentence detection, NER, document categorization and sentiment training model for extraction the value of the parameter. These parameter value helpful for extracting the NLP Text meaning. The response correctness also increases whenever anyone parameter is extracted correctly. The maximum entropy approach is used for making statistical modeling. The training data are taken from the heterogeneous source.

Keywords Sentence detection · NER · POS tagging · Sentiment analyzer · Maximum entropy

A. K. Negi (✉) · S. I. Hassan
Jamia Hamdard University, New Delhi, India
e-mail: anilnegi06@gmail.com

S. I. Hassan
e-mail: s.imtiyaz@gmail.com

1 Introduction

Huge call center load on the human agent in call center try to introduce an NLP-based call center agent who automatically initiates a conversation with the user. We should understand his problem and give him, either a solution or generate a ticket for him. This approach is very important for call center environment because 24 h human involvement with the same alertness and efficiency is not possible. It is observed there are some hours when the human agent is either not available for the user or he has difficulties to take a huge load period of working hours. There is repetitive work nature if we increase the no. of human agents to take the load during the peak working period then the call center service cost will be increased which makes our service infrastructure infeasible. Our call center overcomes this problem when we develop a machine learning NLP-based call center agent for it. This talks with the user just like a human being. The main challenge is learning. How our agent understands the user input. The learning to Chabot about our environment is a very crucial step. How can we use the Apache open NLP Engine for information extraction and Datum box Framework for sentiment extraction from user input? Both Apache Open NLP information extraction and Datum box sentiment extraction used to develop a Chabot agent? the objective is developing an agent which is near to human being understanding. The accuracy of parameter extraction depends on the current knowledge of the parameter. If less knowledge then error-prone response so identify the multiple parameters and implement the ways for extraction of them which increase the efficiency of call center agent.

Chatbot which provides email service for the end user. The user-facing login issue in an email.

User says: I am unable to login in an email.

$$\text{Entities} = \{\text{email}, \text{login}\} \quad (1)$$

$$\text{Category} = \{\text{emaillogin}\} \quad (2)$$

$$\text{Sentiment} = \{\text{negative}\} \quad (3)$$

`Parameters_list1 = {sentence boundary, ner(named entity recognition), dc(document categorization)}`

`Parameter_list2 = {sen}`

If `p(ner = {'email', 'login'}) || p(dc={'emaillogin'})` and `p(negative_sentiment)` then
Call center agent generate the response;

Ok, you have login issue in email, are you agree yes/no.

This approach is used to make a more robust call center agent because if agent succeeds to extract any one parameter, then machine learning agent user interaction becomes successful.

1.1 Maximum Entropy Overview

In a service model which deliver the services like email, network, VPN, etc. There is a call center interface required where user come and interact with machine learning chatbot for registering his complaint.

Response:

$R = \{$

Ok, you have login issue in Email (r1),

Ok, you have new account issue in Email (r2),

Ok, you have forgotten the username or password issue in Email (r3),

Ok, you have relay issue in Email (r4)

$\}$

The user input translated into the corresponding response of email. The conversion depends on the context such as login, new account, forgot username or password, Relay Issue, etc. The probability distribution of email for different context or issues is

$$\begin{aligned} P(\text{email, login}) + p(\text{email, new account issue}) \\ + p(\text{email, forgot username or password issue}) + p(\text{email, relay issue}) = 1 \end{aligned} \quad (1)$$

The initial probability distribution for email complain module is:

$$P(\text{email, login}) = 1/4$$

$$p(\text{email, new account issue}) = 1/4$$

$$p(\text{email, forgot username or password issue}) = 1/4$$

$$p(\text{email, relay issue}) = 1/4$$

The expert sample data is observed and found that 60% time the email complaining about login and forget username or password then as per expert sample data the probability will be re-assigned:

$$P(\text{email, login}) + p(\text{email, forgot username or password issue}) = 6/10 \quad (2)$$

From (1) and (2)

$$\begin{aligned} (\text{email, new account issue}) + p(\text{email, forgot username or password issue}) &= 1 - 6/10 \\ &= 4/10 \end{aligned}$$

So each has the following probability:

$$P(\text{email, login}) = 6/20 = 3/10$$

$$p(\text{email, new account issue}) = 6/20 = 3/10$$

$$p(\text{email, forgot username or password issue}) = 4/20 = 1/5$$

$$p(\text{email, relay issue}) = 4/20 = 1/5$$

we are trying to find out the most suitable model which satisfy both constraint (1) and (2).

1.2 Modeling of Maximum Entropy

For the email example just considered, the process generates a response for user input regarding email problem into r_i which is a member of R .

The response is influenced by some contextual information as

Context = {login, new account issue, forgot username or password issue, relay issue}

The main task is to construct a model which stochastic model that accurately represents the behavior of the random process such a model is a method of estimating the conditional probability, given a context x , the process generates the response y .

$p(y|x)$: the probability of y in the model when the context is x .

$P(y|x)$: entire conditional probability distribution provided by the model.

$p(y|x)$ is just a member of P . P is the set of all conditional probability.

1.3 Training Data Set

To study the process, we observe the behavior of a random process for some time.

Collecting a large set of sample data $(x_1, y_1), (x_2, y_2) \dots \dots \dots (x_n, y_n)$ in this example we take a set of sample data for email complain have context information login in each sample. We consider a case where the training sample data generated by a human expert which represent a number of contexts containing email and asked for a good translation each.

Summarize the Training sample data in term of empirical probability distribution p_{-} Defined by,

$$p_{-}(x,y) = 1/N * (\text{Number of Time } (x,y) \text{ occur in the Sample})$$

The (x,y) is occurring in all the sample, not in all or few samples [1].

1.4 NER (*Named Entity Recognition*)

This is also known as Entity Identification or Entity Extraction or Entity Chunking [2]. This is basically used to extract the Entities such as a person, location, organization, domain, subdomain, etc. from the given information such as

I am unable to login in an Email.

The highlighted part of the sentence is an entity. NER used to extract this information.

There are different areas where the NER has used such as NER Module is used in Azure Machine Learning Studio for extraction of person, organization, and locations. NER is an important area of Machine Learning and NLP (Natural Language Processing). This is used to find out the answer several real-world problems: If anyone dealing with tweet then finds out the tweet has the name of the person.

Does tweet has the name of the location.

Does tweet has the organization.

Other social media data or other information has any entity or not.

Three kinds of Entities are identified by a NER Module. These entities are person, organization, or location [3]. Another area where NER is used is text razor which extracts different entities from several sources of data such as Wikipedia, DBpedia, wiki data etc. These entities saved in a database which is useful to create a dictionary of millions of Entities. Peoples, organization or location are identified by the statistical tagging. We also build a deep contextual understanding of entities in our document. We update the knowledge about entities in our module to update the score of each entity in our document [4].

The information extraction is an area of NLP; NER is also a part of information Extraction technique. This is a predefined concept which is basically used for information extraction. NER is a component of IE in natural language processing. This is basically used to find out the names, person, and location from a piece of information. Manually marking of entities is a very time-consuming process. The recent research is focused on the empirical model, which is built by the training of a supervised model. In near future, this model is used to extract the entities from the given information. I am dealing with the information extraction in call center environment so make a supervised training model using call center domain information. This call center NER Model is used to extract all domain, subdomain, name, location, etc. from the information which is very helpful for generating the appropriate conversational response.

We use supervised learning for it. Label the entities in the information and give this information to our training model generator which generate the NER Training model generator [5]. The focus area of NLP applications is web mining, sentiment analysis, machine translation. This is very important in machine learning, semantic analysis, and information extraction. In social media such as Twitter or Facebook, this becomes an important model for information extraction [6].

1.5 Sentence Detection Model

A sentence is a group of the word which has some meaningful sense. There are several NLP Operations which has been performed on a sentence such as part-of-speech (POS) tagging, machine translation, etc. [7]. Sentence is a terminated by Period (.) sign but there it may differ from a different environment. If we are taking any special case for sentence detection then we create a custom training model, which understands my environmental sentence definition. The sentence information has extracted from user input by consulting with custom training model. The Apache Open NLP provides several sentence detection APIs. We create a text file which has huge no. of a sentence. The training model is a statistical model which learn by provided data set about the sentence boundaries. When the user enters the text then extract the sentence boundary as per supervised learning. Mandarin question sentence detection is a new topic in the field of NLP. It has twofold importance. First, the computer and human dialogue system where they both understand each other and response back and second are the systems of punctuation processing. This type of system required the system knowledge and the queries regarding this environment which is helpful for the system to establish the dialogue between the computer and human being [8]. When we are dealing with Automatically Processing of documents such as summarization, translation, etc., of the Sentence.

Detection is the procedure which is used for that purpose so correctly Identifying the Sentence boundary is an important task in NLP. We can use two approaches to find out either it is a sentence or not. First the fixed rules and second is a Machine Learning Approach to find out it. The first approach will use the fixed rules of Language, it detects the boundary by finding the punctuation and what is the context in which it is used. Then it approaches the Regular Expression Module to make a decision, either as per Language rules, it is a sentence or not [9]. This is a very important approach which is used to find out the sentence boundaries in different languages and also find out the relevant sentences from multilingual languages and summaries those sentences and question-answering. There are so many approaches proposed for establishing the sentence relevancy. Word matching and thesaurus are the approaches which have been adopted for those sentences when two sentences touch the same topic. This is basically focused on the relevancy of sentences in the multilingual Languages [10]. Sentence is an essential and very important block of Natural Language Processing. Syntax parsing, tokenization, chunking, etc. depends on the correct sentence boundary detection. Generally, a sentence is ended with a period (.) but used for abbreviation also, this approach has failed in case of the following:

1. I am living at U.S. where I have been staying since March 2014.

This is an error-prone approach but the output Quality of system is high. The error possibility generally 5–10%.

Rule-based Approach: This approach is also implemented in Apache open NLP. Which have to implement the language rules for detecting the Sentence Boundaries.

The sentence is given to the RE Model which split the text into sentences as per rules [11].

We can also use supervised learning. This approach required a huge set of training data which is used for learning purpose. A machine learning training model is generated for detecting the correct boundary of the sentence.

The sentence boundary detection in the text which is generated due to spontaneous speech is very difficult. The available information is segmented, always this is not problematic to find out the sentence boundaries but in some cases, it becomes a panic issue in Natural Language Processing. Sentence boundary is an unscripted speech which is labeled in a sentence either manually or automatically. Suppose there are five speakers who are speaking a text, the speech is unscripted. How the sentence boundaries are detected [12].

1.6 POS Tagging

This is a very basic problem which tags each word by its grammatical context in the document. It assigns each word a proper morphosyntactic tag to each word in the context the word appears. This is very useful for text to speech system, preprocessing steps for syntactic parsing and corpus linguistics, etc. [13].

I am Suresh Sharma

The pos tagging is:

I_ am_ Suresh sharma_NN

This tagging assigns to each word as per the Penn Treebank corpus.

The POS tagging is very important in the information retrieval, text to speech and automatic translation, etc., the supervised learning for POS Tagging is very time-consuming and expensive approach. This approach is based on the manual tagging of words. This approach getting the high quality of output with respect to unsupervised learning. There are many words in the Arabic language which have ambiguous meaning. This problem is solved by creating a statistical model for our environment. This situation will consult with this statistical model to take the decision regarding POS Tagging of ambiguous words. If the tagging process is better then the output quality is also high. The idea is tokenized the Arabic language words and make the POS Tagging for those words using a conditional random field approach. This approach gets knowledge from the field and create a model based on that knowledge and tag the words as per that knowledge [14].

Molecular biology also uses the POS Tagging approach to find out the tagging for words. This system which is used for its purpose, there are three approaches used: molecular-biology names detector based on rules, handler of an unknown word, and

a tagger based on Hidden Markov model which is used to tag the corpus with the grammatical and molecular biology tags.

There is a large no of molecular biology data are freely available on the internet for research purpose [15].

The approach for the Arabic language, which uses the Hidden Markov Model has introduced by Selçuk Köprü. This paper explores different parameters to improve the baseline system which is used for the Arabic pos tagging. This approach uses the real-life application for tagging and it gets approximate 95.57% accuracy. The pos tagger is widely used in many real-life applications such as Machine Translation, etc. The accuracy of POS tagger has affected the quality and efficiency of the overall system. This is basically used to label the words with POS Tag and other linguistic information. The sentence is tokenized before tagging the word. This research has used approaches for tagging: Rule-based tagging and stochastic tagging. The rule-based approach process a set of rules for pos tagging but the stochastic tagging approach calculate the probability of words pos tagging in a huge data set. Here we analysis different aspect of Arabic POS Tagging and represent it in an HMM [16].

2 Literature Review

The System Required a Machine Learning Call Center Agent. The Literature Reviewed to find out the Different Parameters and Dimension for modeling The Sample Training Data and Extracting The Information and sentiment introduced in 2000, which tells us the POS Tagging using Machine Learning. [11] introduced in 2014 the Tokenization and POS Tagging using the morphological analyzer. [12] introduced in 2011 which find out a solution for POS Tagging in molecular Biology Scientific Abstract using morphological and contextual statistical information. [13] was introduced in 2000 for POS Tagging using the Machine Learning Approach. [14] was introduce in—For improving the Tokenization and POS Tagging in Arabic. [15] was introduced in 2004 for POS Tagging in molecular Biology Scientific Abstract using Morphological and Contextual Statistical Information. [16] was to introduce an efficient POS Tagging approach for Arabic in 2011. [17] introduced a new method of virus detection which is based on the maximum entropy approach.

2.1 Problem Statement

In the Service-oriented System where many service components are deployed. Those components are ready for use. There are several problems occur in day-to-day operations. The user either want an interface where can register his problem or contact for help. In the traditional approach, we establish a call center and a human agent sit there. User calls there and interacts with him for taking help regarding his service model.

The Research describes how we deploy a machine Learning agent, here introduce the steps for extracting the information for getting the semantics of data. The POS Tagging and entities are extracted from the user data to get the data semantic.

The main challenges for developing an English based interacting system are

1. How can understand the user input?
2. What can information extract for better understanding?
3. Same problem described by several users in a different manner.
4. We are generally talking about an entity but which?
5. How can extract the entities?
6. Entity Meaning depends on the context.
7. How my model extracts the entity meaning in different context.
8. The POS Tagging for my service environment.
9. How find out the sentence boundaries in different service environment?
10. Proper class identification using the maximum entropy approach.
11. Which tool can be used for implementation?
12. How can analysis the semantic of data?
13. There are several challenges in making a machine learning model. The model takes the user input and generate most appropriate response.

The other challenges are

1. Analyze the Accuracy of User input.
2. How can the user understand his boundary in that system?
3. If the user crosses his limit then how to handle it.

These challenges have been overcome by making a better understanding of the environment where the agent will be deployed.

Each environment has some entities and process. The research process understands the involved entities and processes. The user involvement in the environment for getting the service regarding those entities. The user gets the services or makes an inquiry regarding those entities.

The Research Process considers an environment and analyzes the sample data for extracting the knowledge and develop a call center agent development template.

The environment has deployed several services such as Email, Network, VPN, etc. There are the following steps to overcome these problems of the Environment.:

Understand the environment. Consider all the entities involved in this environment. The research paper describes, how can extract the entities from user input? Analyze the collected data for this environment. A deep study is required about entities, sentence, and semantics about the service model. The data has classified so that the appropriate path of conversation will be identified. How can we define the Conversation Boundaries of our system? How can the conversation handle if the user going out of track?

3 Proposed Solution

The proposed solution is focused on the path to translate the user input into the appropriate response. The Research paper described the Implementation of Model using the Apache Open NLP. The Apache open NLP has several APIs. The Conversation is initiated on the basis of the following queries:

User talking about which service? Identify the Service.

Service Related to which Entity? Identify the Entity.

Identify sources to collect the previous sample data for Training Purpose and the Training data and our requirement for making a logical conversational session so that conversion reached some logical endpoint. Suppose we are considering a model which replace a call center human agent in an environment where system providing services for email, VPN, cloud, etc. The sample data which have the knowledge about sentence boundaries and entities are save into a text file. This text file has the labeling of sentence boundaries and entities. The text file is given to the Apache Open NLP APIs which generate the training model for making a decision regarding to extract the sentence or entities from the user input. The other dimensions which have been explored by us are document classifier etc. All the data of call center is put here with category and document. Such as if user-facing the Emaillogin issue then

Category = Email login

Issue is: login

The text file which given to the document categorization model has:

Category sentence

Emaillogin I have email login issue.

When the user talks with our agent it detect the category of document. This categorization helps the agent for making a logical dialogue between agent and user.

3.1 Sentence Detection Model Implementation

Steps:

The implementation is based on the Apache open NLP API. It provides the API to construct a stochastic model for a set of sample data [18].

Steps to create a sentence detection model for a set of sample data:

There are two methods:

Method 1:

Step 1: Download the sentence detection model from Apache open NLP sentence detection model.

Step 2: Enter a dialogue with the sentence detection model such as

User says: I have email. I have login issue in email.

Output:

Sentence 1: I have an email.

Sentence 2: I have a login issue in email.

Method 2:

Step 1: Collect all the call center sample data.

Step 2: This data saved in a text file, one sentence per line in sentence.txt. Which have data as

I have an email. I have login issue in email.

Step 3: This file given to the training APIs of Apache open NLP.

Step 4: The generated Training model is saved in a file named “en-sent_custom.bin”.

Step 5: Now

User says: I have email. I have login issue in email.

Output:

Sentence 1: I have an email. I have login issue in email.

This output is based on our custom training model.

3.2 POS Tagger Model Implementation Steps

The Implementation for POS tagging use the Apache Open Natural Language Processing API's, the steps which used for its implementation is described below [18].

Step 1: Download the POS Tagger Training model.

Step 2: Take the user input and give to the sentence detection model. Each sentence pass in the tokenization model which split the sentence into the tokens.

User says: I have login issue. I have an email account.

User input pass to sentence detection model which give the following output.

Output:

Sentence 1: I have a login issue.
Sentence 2: I have an email account.

The sentence passes to the tokenization model and gets the tokens.

Step 3: each token pass to the POS Tagger model for labeling each word as per Penn Treebank. This POS Tagging saved into a table for a further environmental study of word context in the system environment.

Tokens: { 'I', 'have', 'login', 'issue' }

These token pass to the POS Tagger model which labeled these tokens by Penn Treebank.

3.3 NER (*Named Entity Recognition*) Model Implementation

The Named Entity is extracted from user input using the Apache Open Natural Language Processing API. The implementation steps are described below. If we use the existing model for NER Extraction then the output is not as per our environmental requirement such as

I have an email issue.

The email is an entity, but if give user input to Apache Open NLP API's with existing training model then it fails to extract the email entity so we collect our training data set from the expert system. Create a custom training model for NER. This custom model used to extract the entities [18].

Step 1: Collect the sample data from the call center.

Step 2: Analyze the data and labeled the entities in the data.

I am unable to login in an email.

Entity labeling:

I am unable to <START:context> login <END> in <START:area> email <END>.

Step 3: The entity labeled data saved in a text file [19].

Step 4: The text file pass to the Apache Open NLP APIs to generate an entity extraction model.

Step 5: When the user enters the input then this model extracts the entity values for our system parameters.

3.4 Document Categorization Implementation

The documents are categorized as per the area of problem and context. The documents categorization will be done using the Apache open NLP API. The sample data have been collected from the expert system. This sample data are given to Apache open NLP APIs which generate the document categorization model [18].

Document categorization is a requirement based approach so this has not a pre-trained model. We collect the data from the call center. Each sentence labeled with the category. When the user established a dialogue with the agent. The agent takes the input and consults with the document training model. This training model gives the document category. This category guide our agent for response generation.

Generate the Document Generation training model:

The system generates the training model using Apache open NLP document categorization APIs.

Steps for document categorization:

Step 1: Collect the sample data.

Step 2: Save the data in a text file: document-categorization.txt

Step 3: Pass the text file to the Apache open NLP APIs.

3.5 Sentiment Analyzer

The sentiment of user input will be extracted using the datum box machine learning framework API. There are three types of user sentiment are extracted such as “positive”, “negative”, and “neutral” [20].

Steps for sentiment Analyzer;

The Datum box framework used for sentiment analysis;

Step 1: Write a program which is used for sentiment Analyzer.

Step 2: The sentence send to the sentiment analyzer model which return the sentiment of a sentence like “positive”, “negative” or “neutral”.

Step 3: if a sentence has entities and negative sentiment. This sentence has the problem and area of a user problem.

3.6 Call Center Agent Implementation

This algorithm is implemented in Java, which uses the above model for sentence detection, Named Entity Extraction, and document categorization using Apache Open NLP [18] and sentiment analyzer using datum box api [20].

Algorithm:

Step 1: Write a Program which accepts the user input.

Step 2: User input passed to the sentence detector module which returns an array of sentences.

Step 3: these sentences tokenized to find out the POS Tagging for the user input. This tagging is saved into a database for analysis each word tagging context in our call center environment.

Step 4: Each Sentence passes through a NER Program model. This model extracts the entities and the value of the parameter to establishing a meaningful conversation with the user.

Step 5: If the same sentence has a domain, subdomain, and problem. This means, this subdomain of the domain has the extracted problem.

Step 6: The input also passed through Document categorization module.

Example: emaillogin I am unable to login in email.

If category==“emaillogin” and domain==“email” and subdomain==“login” and problem==“unable” and sentiment==“negative” then

Response: Ok, you have the problem with login in an email. Are you generate a ticket for your problem, yes or no?

4 Result and Discussion

The paper describes the way to develop a call center agent which extracts different parameters to find out the sentiment, meaning of user input, which assists our agent to generate the most appropriate response and gives a human being interaction feeling when call center agent respond back. Description of results when the user interacts with the machine learning model:

Sentence detection Results:

User input: I have email. I am login problem.

Output:

I have an email.

I am login problem.

Now generate a custom training model.

User input: email is not working.

Entities: nothing.

Now put the sentence with label entities in text file like

<START: domain> Email <END> is not working. The custom training model is generated for entity identification.

User input: email is not working.

Entities: email (Tables 1 and 2).

Training sample data: one sentence in one line.

I have an email. I am login problem (both sentence this written in a single line).

Save in a text file and generate the statistical training model. Now call center agent behavior.

Show the existing model responses and custom model with training sample with the response (Tables 3 and 4):

POS Tagger:

Use the apache existing pos tagger training model for pos tagging in user input. The word and its corresponding tagging saved in a database for further analyzing the entity context.

Table 1 Entity extraction with the existing model

User input	Existing model response
I am unable to login in Email	Nothing
My email login is not running	Nothing

Table 2 Entity extraction with custom model

User input	Custom model	Training sample data for custom model
I am unable to login in Email	Area : email [probability = 0.4584272427134401]	I am unable to login in <START:area> Email <END>
My email login is not running	Area : email [probability = 0.4626091318397202] context : login [probability = 0.45004963469780435]	My <START:area> email <END> <START:context> login <END> is not running

Table 3 Sentence detection with the existing model

User input	Existing model response	
I have login issue. Unable to login	Sentance1	I have login issue
	Sentance2	Unable to login
I am unable to login. I have a broadband connection	Sentance1	I am unable to login
	Sentance2	I have a broadband connection

Table 4 Sentence detection with custom model

User input	Custom model		Training sample data for custom model
I have login issue. Unable to login	Sentance1	I have login issue. Unable to login	I have login issue. Unable to login
I am unable to login. I have an broadband connection	Sentance2	I am unable to login. I have an broadband connection	I am unable to login. I have an broadband connection

Document Classifier:

The classifier is a custom requirement based model. We save the sentence in a text file with the category. When the user enters the input then this model returns the category such as Document classifier training model generator take a text file which has sentences such as

Category: Emaillogin

Sentence: I am unable to login in an email.

User says: I am unable to login in an email.

The sentence in the training model:

Emaillogin I am unable to login in an email [19] (Table 5).

Sentiment Analyzer:

This model is implemented using datum box API's. User sentence split into sentences. Each sentence sends to the sentiment analyzer. If the sentiment is negative means user have some service-related problem [20].

Table 5 Category classifier model

User input	Existing model response	Training sample data for custom model
I am unable to login in email	Emaillogin	Emaillogin I am unable to login in email

User input: I am unable to login in an email.

Sentiment: negative.

5 Conclusion

There are many fields in computer science, where automation is not possible until and unless the model is not understanding the human being language. The process decision is based on the meaning of user interaction. That's why we are not able to develop a machine model which interacts with a human being and understand his requirement, make a response as per that. The NLP provides the capability to a machine to extract the communication meaning. This meaning used to take the appropriate decision for generating the best possible response which fulfil the user requirement. There are several methodologies as NER, POS Tagging, and sentiment analysis used to get the correct meaning from the human interaction. This information extraction mechanism used to categories the user requirement and attach them with the existing category using the maximum entropy approach. Each category has a kind of response for users. What are the categories? how the particular category behave? and what is the outcome of a particular category is based on the previous knowledge and our existing model, is trained for those using the supervised training or trained the model using different information extraction algorithms.

6 Future Work

This is a field where there is a great diversity of user interaction. The challenge is to capture this diversity, to get the quality output so that interaction between the user and machine learning agent reach at the most possible logical end. Is there any new parameter which explored to improve the output and improve the information extraction methodologies to get better extraction results?

References

1. A Maximum Entropy Approach to Natural Language Processing, <https://aclweb.org/anthology/JJ96/J96-1002.pdf>
2. Named-Entity_Recognition, https://en.wikipedia.org/wiki/Named-entity_recognition
3. Named-Entity_Recognition, <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/named-entity-recognition>
4. Named-Entity_Recognition, https://www.textrazor.com/named_entity_recognition
5. N. Kanya, T. Ravi, Modelings and techniques in named entity recognition-an information extraction task, 23 Jan 2014

6. G. Celikkaya, D. Torunoglu, G. Eryigit, Entity recognition on real data: a preliminary investigation for Turkish, 27 Jan 2014
7. F. Wong, S. Chao, iSentenizer: an incremental sentence boundary classifier, 30 Sept 2010
8. P.-J. Yeh, S.-M. Yuan, *Mandarin Question Sentence Detection: A Preliminary Study* (Springer, Berlin, Heidelberg, 2003)
9. C.N. Silla Jr., C.A.A. Kaestner, *An Analysis of Sentence Boundary Detection Systems for English and Portuguese Documents* (Springer, Berlin, Heidelberg, 2004)
10. M.-H. Hsu, M.-F. Tsai, H.-H. Chen, *Multilingual Relevant Sentence Detection using Reference Corpus* (Springer, Berlin, Heidelberg, 2005)
11. F. Šarić, J. Šnajder, B.D. Bašić, *Optimizing Sentence Boundary Detection for Croatian* (Springer, Berlin, Heidelberg, 2012)
12. A. Stepikhov, *Resolving Ambiguities in Sentence Boundary Detection in Russian Spontaneous Speech* (Springer, Berlin, Heidelberg, 2013)
13. L. Márquez, L. Padró, H. Rodríguez, *Machine Learning Approach to POS Tagging* (Kluwer Academic Publishers, 2000)
14. M.N. Nawar, *Improving Arabic Tokenization and POS Tagging using Morphological Analyzer* (Springer, Switzerland, 2014)
15. G. Dimitris, D. Evangelos, *Part-of-Speech Tagging in Molecular Biology Scientific Abstracts Using Morphological and Contextual Statistical Information* (Springer, Berlin, Heidelberg, 2004)
16. S. Köprü, *An Efficient Part-of-Speech Tagger for Arabic* (Springer, Berlin, Heidelberg, 2011)
17. N.T. Nguyen, V.H. Pham, B.C. Le, D.T. Le, T.H. Van Le, *New Method of Virus Detection Based on Maximum Entropy Model* (Springer, Switzerland, 2015)
18. The Apache OpenNLP Library Is a Machine Learning Based Toolkit for the Processing of Natural Language Text, <https://opennlp.apache.org/>
19. <http://opennlp.sourceforge.net/models-1.5/>
20. Datumbox Machine Learning Framework, <http://www.datumbox.com/machine-learning-framework/>
21. B. Kakheshan, S.I. Hassan, Assessment of accuracy enhancement of back propagation algorithm by training the model using deep learning. Orient. J. Comput. Sci. Technol. **10**(2), 298–304 (2017), <http://dx.doi.org/10.13005/oj cst/10.02.07>. ISSN: 0974-6471, Online ISSN: 2320-8481

Analysis of Energy Consumption in Dynamic Mobile Ad Hoc Networks



Indrani Das, Rabindra Nath Shaw and Sanjoy Das

Abstract Energy is a very crucial parameter in Mobile Ad Hoc Networks since mobile nodes are operated with this scarce resource. If nodes battery is drained, then ongoing transmission in the network is disrupted and discontinued. Mobile nodes can communicate with each other directly or multi-hop fashion. A mobile node uses battery power while working in various modes, i.e., transmitting, receiving, idle, and sleep. Routing protocols, MAC layer, and other network layers exchanges various control packets for executing their task. Overall, performance of the network indirectly depends on the battery power. So, wisely utilizing battery of individual nodes may significantly increase the network performance and lifetime. We have studied the energy utilization of individual node in the network. We have used AODV routing protocol for the analysis of individual node energy utilization in different mode of their operation. Finally, we observed that in receiving mode nodes average energy consumption is higher as compared to transmit mode operation.

Keywords Mobile nodes · AODV · Energy · Battery power · Lifetime · MANET

1 Introduction

The energy of mobile nodes is conserved in Mobile Ad hoc Networks (MANETs) with the help of energy models. This is a self-configured and spontaneously formed the network. Mobile nodes in this network do not follow any fixed infrastructure. Mobile nodes are very frequently joined and leave the network and create a void in the network. Mobile nodes are operated through battery power, and this is a

I. Das

Department of Computer Science, Assam University, Silchar, Assam, India
e-mail: indranidas2000@gmail.com

S. Das (✉)

Department of Computer Science, Indira Gandhi National Tribal University, Amarkantak, India
e-mail: sdas.jnu@gmail.com

R. N. Shaw

Galgotia University, Greater Noida, India

very scarce resource in this network. There is no mechanism for instant recharge or sometimes batteries are non-rechargeable in nature [1]. Mobile nodes are moving in the network randomly or follow a certain pattern depends on mobility model. They can communicate with each other directly if nodes are fall in each other transmission range and indirectly when they do not fall in each other direct transmission range. In this case they formed a multi-hop communication pattern, where data packets are traversed from the source node to destination through intermediate nodes. To achieve seamless connectivity in this network depends on their direction, speed of movements of nodes. In case, single node fails from the ongoing communication, then whole path is broken and route discovery process initiate. This process consumes huge amount of battery power and causes significant delay in the network [2].

In this paper, we have used AODV routing protocol for analyzing the power consumption by individual nodes in various activities, so that overall network life can be analyzed. The result analysis done with the help of variable node density shows that with increasing simulation time power consumption of mobile nodes in different activities are also increases. We have shown the power consumption of individual nodes during data transmit and receive mode.

The paper is organized as; the literature review is discussed in Sect. 2. In Sect. 3, briefly discussed methodology and experimental analysis. We have concluded the paper with future work in Sect. 4.

2 Literature Review

There are so many research works in present days focusing on better utilization of energy of mobile node, performance analysis of various routing protocols, a classification of energy-aware routing protocols, etc., are discussed in this section.

In [3], the authors proposed a cost metric while selecting a path this metric include residual battery and traffic load at a node. Overall, the impact of varying CBR in the term of average energy consumption is analyzed in this paper for varying CBR applications. In this paper [4], a trust-based adaptive stable and energy-aware routing protocol is proposed. In this protocol, the route discovery process finds a stable, reliable and more energy-aware path. This protocol gives better results as compared to other routing protocol in the term of average end-to-end delay and packet delivery ratio.

In [5] authors analyzed DSR, AODV, and OLSR routing protocols for energy efficiency for MANET. This work used AODV, DSR and OLSR routing protocols for the analyze energy consumption. Simulation result observation shows AODV consumed more energy from beginning to end of the simulation. As compare to DSR and OLSR, AODV consumes higher energy in the dense network scenario. There is always a trade-off while choosing routing protocols for MANET, wisely choosing will suddenly improve overall network performance as well as network lifetime. In [6], analysis the energy consumption by the different process of AODV routing

protocol. At the time of route discovery process node energy is not considered, due to this packet drop occur, and this leads to link failure in the network. Therefore system reinitiates RREQ message from the source node, leads to more energy consumption. In this paper, authors proposed a modified AODV protocol which does not consider the link failure problem and refrain itself from rebroadcasting of the message again. The simulation result shows there is a significant improvement in energy consumption. In [7], the issue of battery power utilization is addressed. The energy consumption mobile nodes should be minimized to increase the network lifetime. The authors proposed Minimum Power Consumption Routing (MPCR) protocol and compared with other well-known power routing protocol Minimum Battery Cost Routing (MBCR) and Minimum Total Transmission Power Routing (MTPR). The results show that MPCR outperforms than MTPR.

In [8], evaluating the energy efficiency of DSR, DSDV and AODV routing protocols, including application, network and MAC layer operations and mobile node operation mode such as idle, transmit, sleep and receive. The results show that a substantial amount of energy is consumed at MAC layer [8].

In the ad hoc network [2], if a single node fails due to exhausted energy level of the battery, then ongoing communication breaks due to the path break. Due to this, reinitiate the path discovery process activate and leads to unnecessary power consumption and higher delay in delivery of data. Overall network performance is degraded. This problem is addressed in [2] and proposed an Energy-Efficient AODV protocol (EE-AODV). This algorithm comprises of energy survival and energy-saving phase. The algorithm is capable of selecting shortest path and maintains the reliability of the network.

To discover routes between source and destination node should be energy efficient as well as uninterrupted communication support is always desirable. This paper [9] investigates DSR, DSDV and AODV routing protocols for energy efficiency for healthcare environments. The result analysis shows that DSR is better energy efficient has maximum remaining energy than other routing protocols. The node mobility [10] in MANET causes frequent link failure and create void in the network leads to packet loss and disrupted ongoing communication. This work [10] addressed the energy consumption issues of routing protocols. The DSDV, DSR and AODV routing protocols are included for the analysis The results show that DSDV consumes the minimum energy, increase network lifetime. There are many [X10] energy-saving methodologies existing for analyzing energy consumption, which is included transmission power, reception power, etc.. In [11], the authors calculate the energy utilization of node due to flow in MANET. The transmission and reception of the packets costs is analyzed who is belong to a flow. In MANET, collision is very common and this occurs due to the concurrent flows. To predict the collision additional energy is spent by individual nodes and accurate measurement of this.

In MANET [12], network connectivity is maintained with the multi-hop wireless ad hoc network, where nodes communicate to each other in this manner. Nodes participate in ongoing transmission required additional energy while routing the messages. In this paper, the effect of the following areas namely internetworking between a multi-hop network, and the Internet and the energy utilization. The energy

considered as a function of number of gateways and node mobility pattern is discussed [12]. The experimental result shows that due to the increase number of gateways significantly improves energy utilization of nodes in dynamic network scenarios. This also helps in prevent network partition occurs due to the exhausted battery of mobile nodes.

In the paper, [1] author uses the Markov chain for modeling nodes battery discharge and energy consumption in ad hoc network is analyzed. The major goal of the paper is to reduce the energy consumption with the help of selective start or shutdown of components in the network. In [13] authors surveyed and classifies the existing energy-aware routing protocols as transmission power control and load distribution and sleep/power-down mode approach for MANETs. These algorithms minimize energy required to transmit and receive packets. Also, nodes are spent energy while in idle mode and only listening ongoing communication. The main purpose of this survey, facilitate researchers with all kind advantages and disadvantages of existing protocols and helps in design a more energy-efficient routing mechanism. In [14] authors included a model for evaluating the energy consumption behavior of mobile nodes in MANET. Energy-aware performance analysis of DSR, DSR-*np*, and AODV routing protocols are done.

3 Methodology and Experimental Analysis

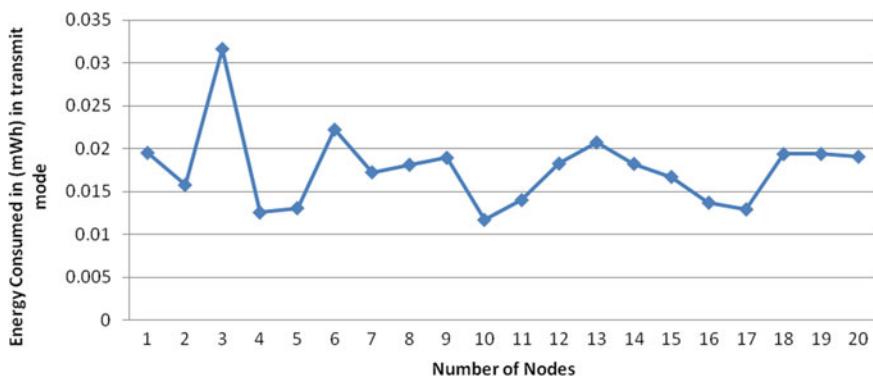
The major aim of our present work is to analyze the battery power consumption of an individual node in the network. To closely monitor transmit and receive modes of operation of mobile nodes we have used AODV routing protocol. Mobile nodes can communicate with each other directly or multi-hop fashion. They are operated in different modes like transmit, receive, idle and sleep mode. For every packet receive or transmit nodes are spent a certain amount of energy. Our analysis focus on nodes performance while working in transmits and receives mode, and how much energy is consumed and left. A dynamic network scenario is considered here. We have varied the simulation time to analyze and validate the energy consumption by AODV routing protocol. The simulation parameters used in our analysis is given in Table 1.

3.1 Energy Consumption in Transmit Mode

In Fig. 1, we have shown the energy consumption of individual nodes for the simulation time 500 s during transmit mode. The minimum energy consumed by node 10 is 0.0117 due to less number of data packets send and less participate in sending other control packets as well. Energy consumption is maximum by node 3 is 0.03162 because this particular node transmits a large number of data and control packets. Energy consumption is varying for individual nodes throughout the simulation time.

Table 1 Simulation parameters

Parameter	Values
Simulation time	500,1000 and 2000 s
Number of nodes	20
Network area	1500×1500
Number of channels	1
Path-loss MODEL	Two ray
Antenna model	Omnidirectional
Radio type	802.11b
Data rate	2 Mbps
Packet reception model	PHY802.11b
Routing protocol	AODV
Energy model	Mica-motes
Mobility model	Random waypoint
Pause time	1 s
Min. speed	0 mps
Max. speed	10 mps

**Fig. 1** Energy consumption (simulation time 500 s.)

In Fig. 2, we have shown energy consumption of individual nodes for the simulation time 1000 s during transmit mode. The minimum energy consumed by node 17 is 0.025 maximum by node 3 is 0.04841. Energy consumption is varying for individual nodes throughout the simulation time.

In Fig. 3, we have shown energy consumption of individual nodes for the simulation time 2000 s during transmit mode. The minimum energy consumed by node 10 is 0.05313 maximum by node 3 is 0.08439. Energy consumption is varying for individual nodes throughout the simulation time.

Mobile nodes working in transmit mode, energy consumption is higher for simulation time 2000 s. Energy consumption is 54% higher than simulation time

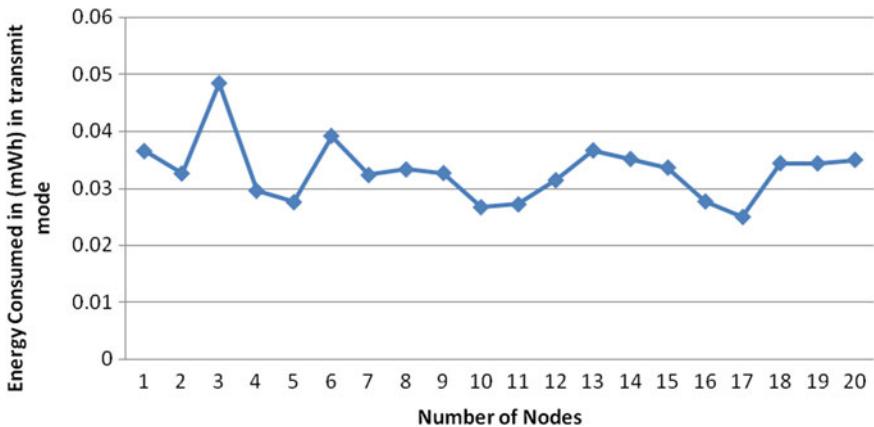


Fig. 2 Energy consumption (simulation time 1000 s.)

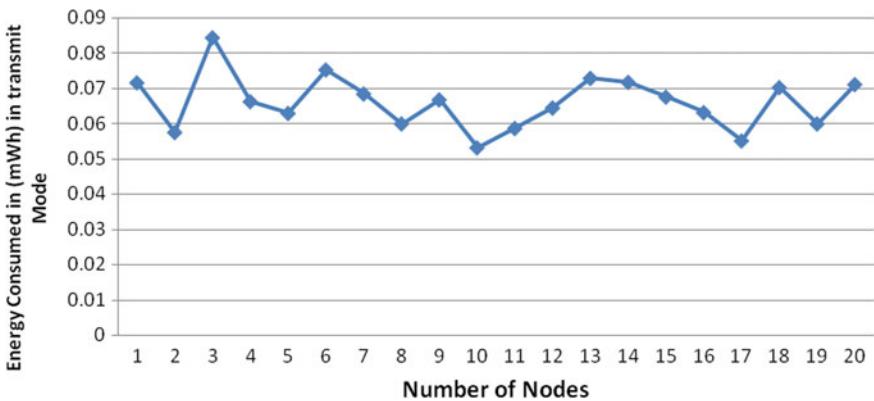


Fig. 3 Energy consumption (simulation time 2000 s.)

1000 s as compared to simulation time 500 s. Also, in simulation time 2000 s consumed approximate 50% higher energy, as compared to energy consumption during simulation time 1000 s.

3.2 Energy Consumption in Receive Mode

In the Fig. 4, we have shown the energy consumption of mobile nodes when working in receive mode. We have varied the simulation time for proper analysis. The result clearly shows that with increasing simulation time consumption on energy is also

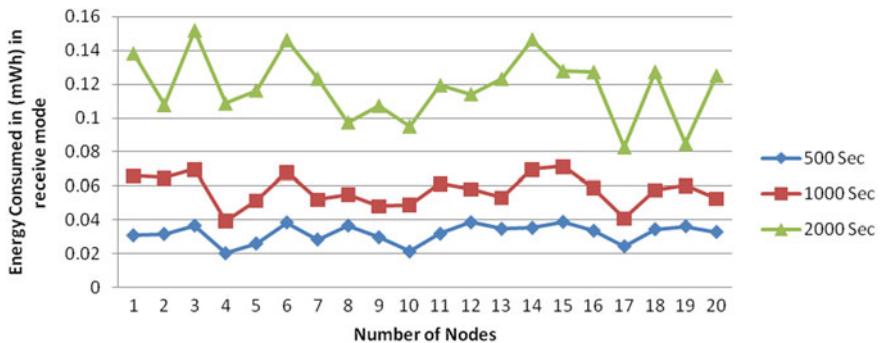


Fig. 4 Energy consumption (receive mode) with varying simulation time

increased. Energy consumption is increased average approximately 56% when simulation time increased from 500 to 1000 s and approximately 48% when simulation time increased from 1000 to 2000 s.

3.3 Cumulative Energy Consumption

In Fig. 5, the comparison of energy consumption both in transmit and receive mode is shown. Different simulation time is considered in analyzing battery power consumption. From Fig. 5, it is clear that during receive mode nodes utilizes their maximum energy. During simulation time 500 s overall nodes consumed 54% higher energy as

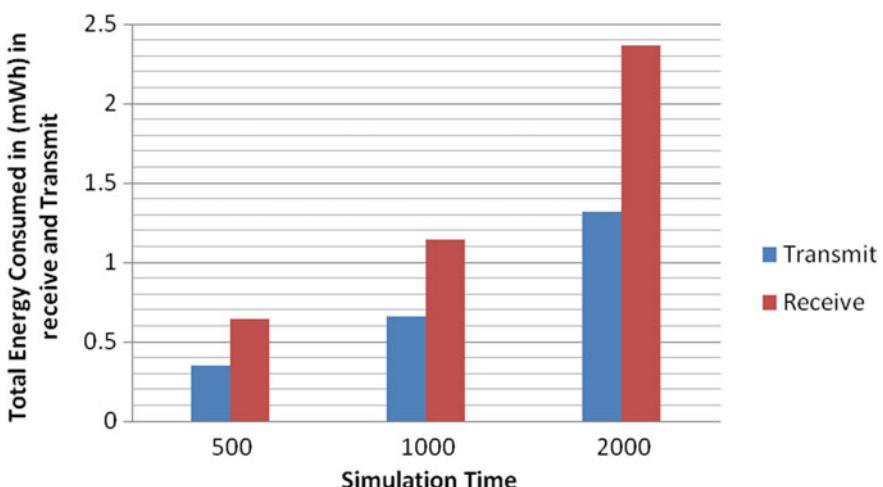


Fig. 5 Total Energy Consumed in (mWh) in receive and Transmit

compared to mode work in transmit mode. Further, during simulation time 1000 and 2000 s nodes are consumed nearly 57 and 55% higher energy in receiving mode as compared to transmit node.

4 Conclusion

In MANET, mobile nodes are operated with so many constraints like limited battery power, dynamic network scenario, limited bandwidth, etc. Our study is to analyze battery power consumption by various network activities i.e. receiving messages, sending messages, sending control and other messages etc. Through simulation we have observed that battery power gradually decreases with increasing simulation time. Mobile nodes are operated in transmit, receive, sleep and idle mode. Energy is dissipated while working in different modes. This study helps in analyzing the lifetime of the overall network. We have shown the individual node energy consumption based on nodes activities. Our result analysis also shows that during receive mode nodes are consumed higher energy as compared to transmit mode of operation. Longer simulation time more energy consumption. So, mobile nodes operate for long time period definitely exhausted their battery power and create disrupted communication in the network.

References

1. M. Heni, A. Bouallegue, R. Bouallegue, Energy consumption model in ad hoc mobile network
2. V.N. Palav, S.R. Bhosale, Energy consumption in MANETs using energy efficient AODV protocol. *Int. J. Adv. Res. Comput. Eng. Technol. (IJARCET)* **3**(4), 1463–1468 (2014)
3. S. Kabbur et al., Impact of CBR traffic on energy consumption in MANET, in *IOP Conference Series: Materials Science and Engineering* (2017)
4. S. Sarkar, R. Datta, An adaptive protocol for stable and energy-aware routing in MANETs. *IETE Tech. Rev.* **34**(4), 353–365 (2017)
5. M.T. Sultan, S.M. Zaki, Evaluation of energy consumption of reactive and proactive routing protocols in MANET. *Int. J. Comput. Netw. Commun. (IJCNC)* **9**(2), 29–38 (2017)
6. A. Tiwari, I. Kaur, Performance evaluation of energy efficient for MANET using AODV routing protocol, in *2017 3rd International Conference on Computational Intelligence & Communication Technology (CICT)*, Ghaziabad (2017), pp. 1–5
7. R. Prasad, P. Shivashankar, Improvement of battery lifetime of mobility devices using efficient routing algorithm. *Asian J. Eng. Technol. Appl.* **1**(1), 13–20 (2017)
8. H. Xiao, D.M. Ibrahim, B. Christianson, Energy consumption in mobile ad hoc networks, in *2014 IEEE Wireless Communications and Networking Conference (WCNC)*, Istanbul (2014), pp. 2599–2604
9. S. Abid, I. Shafi, S. Abid, Energy efficient routing in mobile ad-hoc networks for healthcare environments. *IJCSI Int. J. Comput. Sci. Issues* **10**(1), 497–505 (2013)
10. A. Kumar et al., Performance evaluation of energy consumption in MANET. *Int. J. Comput. Appl.* (0975–8887) **42**(2), 7–11 (2012)

11. G. Allard, P. Minet, D.Q. Nguyen, N. Shrestha, Evaluation of the energy consumption in MANET, in *Ad-Hoc, Mobile, and Wireless Networks. ADHOC-NOW 2006*, ed. by T. Kunz, S.S. Ravi. Lecture Notes in Computer Science, vol. 4104 (Springer, Berlin, Heidelberg)
12. S.S. Rao, K.C.K. Reddy, Energy consumption model in MANET for mobile ad hoc networks. *Int. J. Res. Advent Technol.*, Special Issue, RAECE-2K19, pp. 51–57
13. C. Yu, B. Lee, H. Yong Youn, Energy efficient routing protocols for mobile ad hoc networks. *Wirel. Commun. Mob. Comput.* **3**, 959–973 (2003)
14. L.M. Feeney, An energy consumption model for performance analysis of routing protocols for mobile ad hoc networks. *Mob. Netw. Appl.* **6**, 239–249 (2001)

Improved ITCA Method to Mitigate Network-Layer Attack in MANET



Nilesh R. Marathe and Subhash K. Shinde

Abstract MANET is well known for its inherent feature of on-demand ad hoc establishment of network. This makes MANET a suitable option for many applications like Disaster management, military applications, etc. But the mutual dependency among the nodes make the MANET vulnerable for many attacks. Researchers had proposed many solutions to make the routing in MANET secure, ITCA is one of the proposed examples which tries to identify malicious activity and isolate infected nodes from network through multiple dimensions. The Improved ITCA proposed in this paper tries to make attack detection system real time and trust calculation adaptive to the application-specific parameters. This will reduce the burden over source node, which has been used by most of ACK-based solution for attack identification and isolation, in turn tries to reduce the number of control packet required that optimize the overhead and make the attack detection and isolation process more simple and faster. The Improved ITCA introduces a lightweight real-time option for secured ACK-based approach, whereas adaptive application-specific trust calculation parameters make the system more robust or work efficiently even when percentage of malicious node in the network is high.

Keywords MANET security · Ad-hoc security · Secure AODV · Black hole attack · Trust · ACK-based system

1 Introduction

The self-organized, on-demand, infrastructure-less implementation of network makes the MANET most useful for ad hoc applications. Many emergency services or military-based application finds the MANET a very useful solution. The mutually

N. R. Marathe (✉)
RAIT, Nerul, Navi Mumbai, India
e-mail: nilesh.marathe@rait.ac.in

S. K. Shinde
LTCOE, Kopar Khirane, Navi Mumbai, India

dependent nodes in MANET catches the attacker attention to exploit many vulnerabilities and attacks. The infected malicious nodes are injected in the route establishment to ruin the routing process and reduce the performance of system . While focusing on ad hoc on-demand reactive routing option, AODV is most popular. THE RREQ, RREP and RERR are the three major control packets used in AODV for route establishment. The attacker use mutual dependent nodes structure of MANET to threaten routing process by exploiting the route establishment by advertising spurious RREQ and REEP. In such scenario, making the route establishment process more robust and data transmission more reliable is the major challenge for researchers. The solution proposed many researchers can be bifurcated as IDS (ACK-based solutions) and node reliability checkers (Trust based solutions).

EAACK [1], AACK [2, 3], TWOACK [4] are some of the examples of the ACK-based IDS solutions. These systems make source node overburden by putting responsibility of attack detection and malicious nodes isolation over it. It also delays the process exact infected node isolation and lose lot of packets which reduce the performance of the system. At the same time, it works with the belief that their tracking system gets genuine response from nodes which are all honest. So verification of nodes which act as a monitor for attack detection has to be incorporated in the proposed IDS which demands to have an additional dimension in support of the existing IDS solution to make it more efficient.

The published ITCA [5] method tries to support routing process from multiple dimensions as IDS and Trust to mitigate network-layer attacks but still utilizes the source node for attack detection. The proposed improved ITCA tries to make attack detection process real time where every intermediate node involved in route is responsible for attack detection and make the trust system adaptable to the application-specific parameters. The new improved ITCA is elaborated in detail in this paper in the following sections.

The paper is organized as Sect. 2 gives related work, Problem definition and Proposed improved ITCA are given in Sects. 3 and 4 followed by Sect. 5 conclude the paper.

2 Literature Survey

The EAACK [1] proposed by Shakshuki et al. is an Intrusion Detection System (IDS). It uses three modes, namely, ACK, S-ACK, and MRA. It will detect the malicious link having two nodes declared as malicious using MRA report. So detection of exact node is challenge than link which may have one legitimate node declared as malicious.

Adaptive ACK (AACK) [2, 3] improvement over TWOACK successfully able to reduce the overhead but again fail to reach exact malicious node reach up to malicious link.

The TWOACK [4] scheme asks the every two-hop node in forward direction toward destination and asks to send acknowledgment to the previous two-hop node

in reverse direction in a group of three consecutive nodes. When any of the Two Hop ACK is not received, it declares the group as malicious which will mark legitimate nodes also malicious and affect on availability of nodes for data routing.

Patil has proposed Improved EAACK [6] scheme where it proposes the special mode to find exact malicious node. The last node in path which transfers files successfully considered as trusted node and asks to give feedback using as NACK to source node which then forwards special packet to next suspicious node if not responded then declared as malicious. The system requires more control packet also it delays the node detection which tends to more packet loss and overload the source node.

Swapnil has proposed the generic request reply based mechanism [7, 8], which tracks the suspicious node based on reply of forwarded packet status given by each node in response source node probing to each node and then the node which has suspicious node in its vicinity ping the node with fake RREQ if responded then declare as malicious. The attack detection is not real time so it does delayed detection and also considers all nodes are legitimate and block malicious node.

The ITCA [5] proposed a multidimensional approach to mitigate the network layer attacks. It improves the efficiency of routing process by supporting IDS and Trust-based scheme. But puts a lot of burden on source node also the trust calculation may be multi-attribute based so that additional factors can be considered in measuring the node reliability.

Ullah [9] has enlisted different issues with trust schemes proposed for MANET security. The major advantage of Trust management schemes is it can work in absence of centralized controller for handling internal attacks on routing and forwarding process. It recommends to have trust calculation based on direct and indirect information as well as has to consider the Direct and Indirect trust calculation for concluding the trust on any node. Thus, our proposed system incorporates the suggestions and provides the multi-attribute based trust model.

Cai has proposed an evolutionary self-cooperative trust (ESCT) [10] where every node does the self-detection for direct neighbours as benign or malicious and calculate trust for it. Then it initiates the cooperative detection and asks to share other node experience to decide the node as malicious or genuine. For that, it needs to broadcast hello messages periodically and maintain the a matrix of opinions which will be overhead for mobile nodes so for lightweight applications it will be definitely overhead in terms of traffic get generated, memory and also energy.

Ghone proposed a reputation-based and trust-based module for detection of selfish nodes in MANET [11]. The proposed algorithm consists of multiple modules to select secured path for routing. It has neighbour monitoring module to observe the behaviour of the neighbours then path ranker to rank the path according to the positive or poor occasions. Involving much more attributes for path ranking might help to improve the efficiency of the algorithm. It considers the RREQ and RREP for concluding behaviour of the node which will be much single dimension that will be the limitation of the proposed algorithm.

Cross-layer approach collect the data from MAC and Network layer and tries to attempt black hole attack [12]. If it results in the identification of some suspicious activity then it initiates the malicious node detection process. The proposed method broadcasts the fake request from source node to track and segregate the malicious nodes. The broadcasting of fake request from source node may create a lot of traffic rather if broadcasts for one hop only by specific node which has the suspicious node in his vicinity will control the broadcast traffic. Rather than collecting information then detecting the attempt of black hole attack better to detect it run time and corrective action can be taken immediately to improve the ongoing transaction delivery.

Nachammai [13] had proposed a method, which combines the Cooperative Bait detection approach with RREQ and RREP based reverse tracing technique. The concept of Fake RREQ is also incorporated in this mechanism but by tracing the attack in a reverse way. This is again not real-time detection so once the transaction gets hampered then it initiates the detection process, it is the major limitation of this method.

3 Problem Definition

The main aim of the proposed system is to improve the multidimensional ITCA [5] approach in terms of ACK and Trust-based solutions to reduce control packets overhead and make it more efficient. It deals with some of issues as

- Make the attack detection system real time by involving each intermediate node in the route as monitoring node so that source node is bit relaxed.
- Real-time approach increases the speed of attack detection with less control packet required and avoid bottleneck at source node.
- Involve more attributes in trust calculation that can be re-configured according to application used.

4 Proposed Framework, Improved ITCA

The consolidated multidimensional multi-attribute-based approach is key behind the proposed methods elaborated in this paper. The ITCA [5] put the burden on source node to track the malicious activity and detect the intentionally misbehaving nodes. This will be time consuming and uses additional control packets that increases the overhead. Rather if able to detect malicious activity at real time with each node which observe this misbehaviour of neighbour immediately initiates the malicious node detection process which will improve the efficiency of existing method ITCA [5].

4.1 Improved Attack Detection and Malicious Node Isolation

Attack detection phase

In the existing ACK based methods like ITCA [5] EAACK [1], AACK [2, 3], TWOACK [4] source node probe for the packet forwarding status to each node in the route towards destination node as shown in Fig. 1.

As given in Fig. 1, source node A pings the node B, C and D which take part in the route to reply packet forwarding status. The received reply from intermediate nodes will be measured for declaring the attempt of attack. The nodes which reply forwarding status as zero and the previous node of that node will be broadcast as suspicious nodes. In EAACK [1] source node will trace the exact malicious node whereas in Request Reply method proposed by Swapnil [8] any node which has suspicious node in his vicinity initiates the exact malicious node isolation process by initiating the fake request packet. But in both cases the node which tries to track the exact malicious node may itself be also malicious, this has been taken care ITCA [5] by allowing only forwarders to initiate the isolation process, explain in detail in article published ITCA [5].

The ITCA can be improved by enabling each node participated in the route to track the previous hop node behaviour based on the delay in packet delivery as shown in Fig. 2.

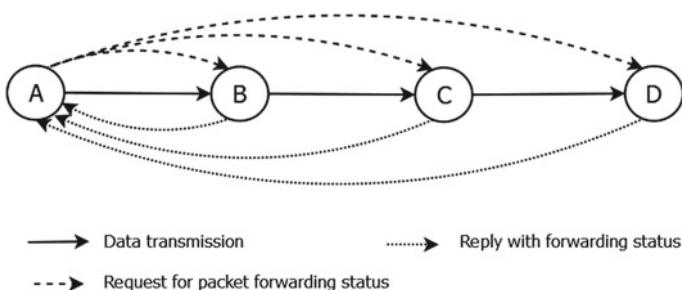


Fig. 1 Attack detection in ACK based methods

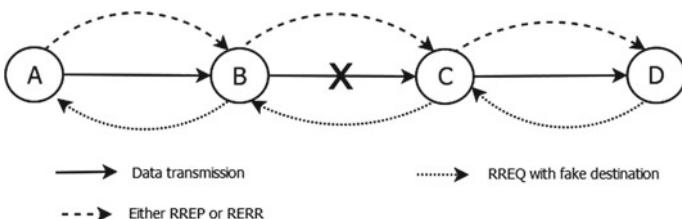


Fig. 2 Attack detection phase improved ITCA

After route discovery when a node forwards the RREP it expects the data to be received within $Data_{Timer}$ duration from the previous hop node. The $Data_{Timer}$ can be calculated using Eq. 1 as given below

$$Data_{Timer} = \begin{cases} 2 * RTT, & \text{Initial value} \\ 2 * \frac{\sum_{k=0}^n (RTT)}{n} & \text{for all future transactions} \end{cases} \quad (1)$$

So whenever packet delivery is delayed and the waiting time goes above average that is the indication of presence of malicious node in the considered route which is responsible for delay or loss of packet as shown in Fig. 2. This will force to initiate the malicious node isolation phase.

Each node in the route keeps track of how many packets forwarded from source to destination. When any one of the nodes involved in the route does not receive data packet it can initiate the malicious node detection and isolation process. So rather than depending on the source node to keep track, every node on the path is responsible for tracking the attack.

Algorithm 1 elaborates the attack detection technique discussed. Attack detection

```

Result: Attack Detection System
initialization;
while Data_Transmission do
    Initiate route discovery process;
    Select the node with role as “forwarder” or “Recommender” to be part of route.>;
    if DataTimer expires then
        If node finds currently established or requested route inactive
        Consider the node as suspicious node and initiate the node detection and isolation
        process.>;
    end
end

```

Algorithm 1: Algorithmic steps for attack detection

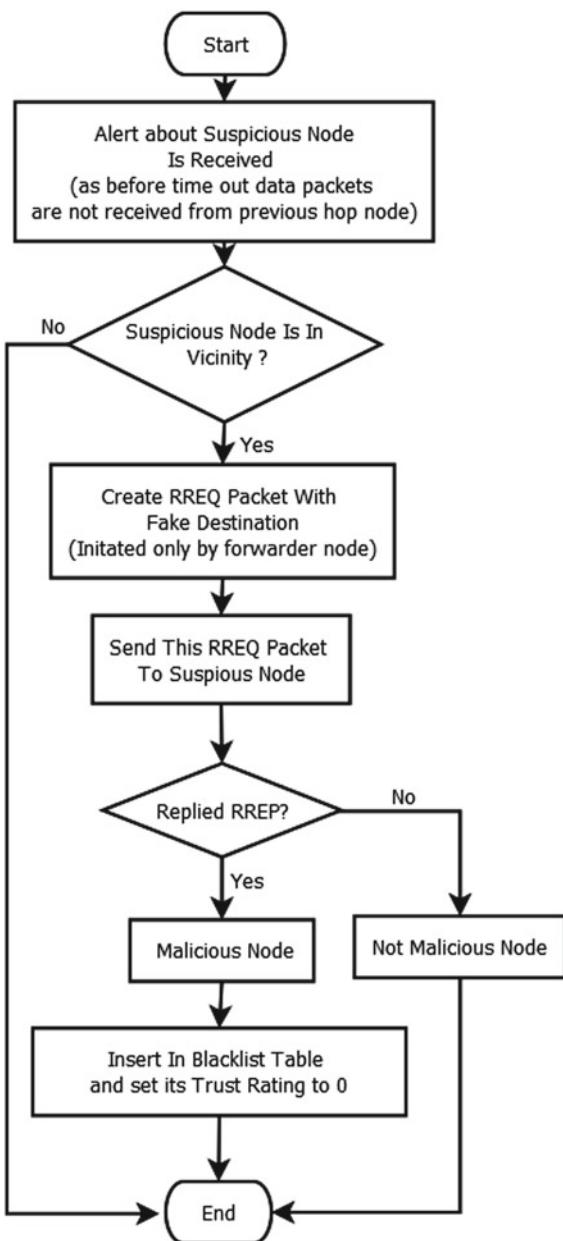
phase has to be supported by system which concludes the exact malicious nodes from the list of suspicious nodes, it is performed by the isolation function explained in next A.

Malicious node detection and isolation Phase

This process of the exact culprit identification is initiated by the intermediate nodes in the route which senses the misbehaviour by previous nodes. So such intermediate node forwards the fake request packet towards the suspicious node if it gets reply then it declares that replied node as inherent malicious node and blacklist it to get isolated from the network route. Flowchart for the same is shown in Fig. 3.

This explains in detail the first dimension of the improved ITCA that is ACK based approach but more smartly designed malicious nodes may not respond to fake request so they may not get tracked by only one dimension, it has to be supported by Trust based approach another pillar of the Improved ITCA explained in next section.

Fig. 3 Flow chart for malicious node detection



4.2 Node Reliability: Trust Value ($trustVal_N$)

Trust value is a major of reliability of nodes which help to track dynamic behaviour of malicious nodes. The trust calculations are possible to make simpler by considering a single attribute but it can be made more stronger by incorporating more attributes for direct trust calculation as given in Eq. 2. The node has categories as Recommender, Forwarder and User. The every node has given chance to showcase its reputation by defining it by default as forwarder that is the trust value equal to 0.4 which updates dynamically as given by Eq. 2.

$$trustVal_N = \begin{cases} 0.4, & \text{By default} \\ trustVal_N + 0.04, & \text{If } IITCA_{Trust} \geq 0.7 \\ trustVal_N - 0.08, & \text{otherwise} \end{cases} \quad (2)$$

where $IITCA_{Trust}$ given by Eq. 3

$$IITCA_{Trust} = \beta * (S_{val}) + \alpha * (Deliverytime_{ratio}) \quad (3)$$

Where, $\beta + \alpha = 1$ are the weights assigned to the attributes S_{val} (4) and $Deliverytime_{ratio}$ (5).

Satisfaction Value:- S_{val}

$$S_{val} = \sum_{k=0}^n (W_k) a_k \quad (4)$$

W_k is the weight assigned to attribute a_k . These attributes are generalized as packet size, number of packets forwarded successfully, etc., based on measures defined for a specific application.

Delivery time ratio:- $Deliverytime_{ratio}$

$$Deliverytime_{ratio} = \frac{Delivery_{Act}}{Delivery_{Exp}} \quad (5)$$

It is the ratio of the ensured delivery time versus actual delivery time of the product.

5 Conclusion

On-demand establishment of networks is the key requirement of many ad hoc based applications like Military based, Disaster management, etc. The new upgrowing field MANET is the best example of mutually dependent node network but it becomes

major vulnerability for many attacks. So the researchers are trying to make the MANET routing robust and secure. The proposed Improved ITCA method is the one step taken to make routing secure from multiple dimensions to mitigate network layer attacks. Since ITCA is implemented successfully using Qualnet the few modifications in existing system suggested in this paper will improve the efficiency by reducing overhead. Here the Improved ITCA tries to boost the speed of attack detection and include the application specific attributes in trust calculation to broaden the scope and make system more generalized.

References

1. E.M. Shakshuki, N. Kang, T.R. Sheltami, EAACK—a secure intrusion-detection system for manets. *IEEE Trans. Ind. Electron.* **60**(3), 1089–1098 (2013)
2. T. Sheltami, A. Al-Roubaiey, E. Shakshuki, A. Mahmoud, Video transmission enhancement in presence of misbehaving nodes in manets. *Multimed. Syst.* **15**(5), 273–282 (2009). <https://doi.org/10.1007/s00530-009-0166-0>
3. D. Sandhiya, K. Sangeetha, R. Latha, Adaptive acknowledgement technique with key exchange mechanism for manet, in *2014 International Conference on Electronics and Communication Systems (ICECS)*, Feb 2014, pp. 1–5
4. K. Balakrishnan, J. Deng, V.K. Varshney, TWOACK: preventing selfishness in mobile ad hoc networks, in *IEEE Wireless Communications and Networking Conference, 2005*, vol. 4, March 2005, pp. 2137–2142
5. N. Marathe, S.K. Shinde, ITCA, an IDS and trust solution collaborated with ACK based approach to mitigate network layer attack on MANET routing. *Wirel. Pers. Commun.* (2019), <https://doi.org/10.1007/s11277-019-06282-5>
6. A. Patil, N. Marathe, P. Padiya, Improved EAACK scheme for detection and isolation of a malicious node in manet, in *2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, Oct 2015, pp. 529–533
7. S. Bhagat, P. Padiya, N. Marathe, A generic request/reply based algorithm for detection of black-hole attack in MANET: simulation result, in *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, vol. 00, July 2017, pp. 1–7, <https://doi.org/10.1109/ICCCNT.2017.8204058>
8. S.P. Bhagat, P. Padiya, N. Marathe, A generic request/reply based algorithm for detection of blackhole attack in MANET, in *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, Aug 2017, pp. 1044–1049
9. Z. Ullah, M.H. Islam, A.A. Khan, Issues with trust management and trust based secure routing in MANET, in *2016 13th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, Jan 2016, pp. 402–408
10. R.J. Cai, X.J. Li, P.H.J. Chong, An evolutionary self-cooperative trust scheme against routing disruptions in manets. *IEEE Transact. Mob. Comput.* **18**(1), 42–55 (2019)
11. M.M. Ghonge, P.M. Jawandhiya, V.M. Thakare, Reputation and trust based selfish node detection system in manets, in *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, Jan 2018, pp. 661–667
12. J.K. Vinayagam, C.H. Balaswamy, K. Soundararajan, Adopting cross layer approach for detecting and segregating malicious nodes in MANET, in *2017 International Conference on Signal Processing and Communication (ICSPC)*, July 2017, pp. 457–461
13. M. Nachammai, N. Radha, Securing data transmission in MANET using an improved cooperative bait detection approach, in *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, Oct 2016, pp. 292–297

Employing Machine Learning Models to Solve Uniform Random 3-SAT



Aditya Atkari^{ID}, Nishant Dhargalkar^{ID} and Hemali Angne^{ID}

Abstract We have employed a chosen set of machine learning models to solve the 3-CNF-SAT problem. Through f1-scores, we obtain how these algorithms perform at solving the problem as a classification task. The implication of this endeavour is exciting given the property of the NP-complete class problems being polynomial-time reducible to each other.

Keywords NP-completeness · SAT · 3-CNF-SAT · Machine learning · Classification · SATZilla features · TensorFlow

1 Introduction

We seek to employ machine learning models to solve the 3-CNF-SAT problem, a version of the satisfiability problem, and deduce how these models perform based on the metric of f1-scores. The dataset used is the uniform random 3-SAT dataset from the SATLIB repository. This uniformity refers to the fact that the variables in each of the instances are distributed evenly in all of the instances of a dataset. This results in the mean of variables participating in clauses to be the same for each instance. The features used are a subset of SATZilla features used in [1]. These are features sets derived from variable–clause bipartite graphs, variable–variable graphs and the ratio of occurrences of positive literals to the negative literals of a variable throughout an instance.

A linear classifier, dense neural network classifier, decision tree classifier, random forest and naive Bayes models were trained on this dataset. The testing results show that the models on this dataset, with the selection of said features, perform underwhelmingly. Amongst them the random forest classifier gives the best f1-score followed by the decision tree classifier.

A. Atkari (✉) · N. Dhargalkar · H. Angne
MCT's Rajiv Gandhi Institute of Technology affiliated to Mumbai University, Andheri West,
Mumbai 400053, Maharashtra, India
e-mail: adityaatkari@gmail.com

The Satisfiability problem, a representative of the NP-Complete class, is important from both theoretical and practical perspectives. This problem occurs in a variety of domains including hardware verification, security protocol analysis, theorem proving, scheduling problems, routing, planning, artificial intelligence as well as digital circuit design. This paper looks to machine learning to find effective ways to solve NP-complete problems by having the satisfiability problem be a representative of the class. Endeavouring further to better the performance of these models to solve SAT will have profound implications on numerous fields where an extension of the problem does very much exist.

2 NP

Decision problems can be classified as P and NP, P problems are deterministic polynomial-time problems, whereas NP problems are nondeterministic polynomial time problems. Decision problems are assigned complexity classes (such as NP) based on the fastest known algorithms. Therefore, decision problems may change classes if faster algorithms are discovered.

It is easy to see that the complexity class P (all problems solvable, deterministically, in polynomial time) is contained in NP (problems where solutions can be verified in polynomial time), because if a problem is solvable in polynomial time then a solution is also verifiable in polynomial time by simply solving the problem. But NP contains many more problems, the hardest of which are called NP-complete problems. An algorithm solving such a problem in polynomial time is also able to solve any other NP-complete problem in polynomial time.

If there is a polynomial-time algorithm for even one of them, then there is a polynomial-time algorithm for all the problems in NP-complete. Because of this, and because dedicated research has failed to find a polynomial algorithm for any NP-complete problem, once a problem has been proven to be NP-complete this is widely regarded as a sign that a polynomial algorithm for this problem is unlikely to exist. However, in practical uses, instead of spending computational resources looking for an optimal solution, a good enough (but potentially suboptimal) solution may often be found in polynomial time. Also, the real-life applications of some problems are easier than their theoretical equivalents.

All NP-complete problems are reducible to each other i.e if a polynomial time solution were to be found for one NP-complete problem then we'd have found a polynomial time solution to all NP-complete problems. The reductions are all of polynomial time. This is also known as polynomial-time reducibility.

3 SAT Problem

We have defined the notion of an NP-complete problem, but up to this point, we have not actually proved that any problem is NP-complete. Once we prove that at least one problem is NP-complete, we can use polynomial-time reducibility as a tool to prove other problems to be NP-complete. Thus, we now focus on demonstrating the existence of an NP-complete problem: the circuit-satisfiability problem.

The aim of the circuit-satisfiability problem is to reduce a sub-circuit which always gives unsatisfiable results. This would help a practitioner to replace the sub-circuit by an off signal. Determining such sub-circuits can be tough hence, circuits are reduced to formulae to further analyze them. The circuit-satisfiability problem is hence reduced to formula-satisfiability.

This problem has the historical honour of being the first problem ever shown to be NP-complete. We formulate the (formula) satisfiability problem in terms of the language SAT as follows: An instance of SAT is a Boolean formula composed of

n Boolean variables: x_1, x_2, \dots, x_n ;

m boolean connectives: any Boolean function with one or two inputs and one output such as \cap (*AND*), \cup (*OR*), \sim (*NOT*), \rightarrow (*implication*), \leftrightarrow (*if and only if*); and parentheses. (Without loss of generality, we assume that there are no redundant parentheses, that is, a formula contains at most one pair of parentheses per Boolean connective.)

Equations in SAT can be of any form and require us to consider multiple cases to handle these forms. The language of these equations can be reduced so as to prevent us from considering too many cases for the equations but still maintaining the NP-complete status of the problem. The problem derived after language restriction is 3-CNF-SAT. 3-CNF-SAT has 3 literals per clause, in Conjunctive Normal Form (that is, conjunction of fundamental disjunctions) and is a satisfiability problem.

A Boolean or propositional Formula can be represented as an AND of ORs.

Example:

$$(\overline{x_1} \cup x_2) \cap (\overline{x_2} \cup x_1)$$

Where:

$\overline{x_1}, x_2, \overline{x_2}$, x_1 are all literals

$c1 : (\overline{x_1} \cup x_2)$, $c2 : (\overline{x_2} \cup x_1)$ are both clauses

A formula is satisfiable if there exists a satisfying assignment.

Boolean expression in 3SAT form:

Expression containing 2 clauses, each clause containing 3 literals:

$$(x_1 \cup x_2 \cup x_3) \cap (x_4 \cup x_5 \cup x_6)$$

$$(x_1 \cup x_2 \cup x_3) \cap (\overline{x_1} \cup x_5 \cup x_2)$$

Expression containing 3 clauses each clause containing 3 literals:

$$(x_1 \cup x_2 \cup x_3) \cap (x_4 \cup x_5 \cup x_6) \cap (x_7 \cup x_8 \cup x_9)$$

$$(x_1 \cup x_2 \cup x_3) \cap (x_4 \cup \overline{x_5} \cup x_6) \cap (\overline{x_1} \cup x_5 \cup x_3)$$

4 Dataset and Features

4.1 Features

Various considerations were made while selecting the features. For instance, the presence of a literal in an equation whose conjugate doesn't appear in the equation increases the chance of the equation being satisfiable. This is because this literal can be assigned a true value (true if its a positive literal and false if its a negative literal) and we can be assured that all clauses it occurs in will be true. Additional features that influence the satisfiability of the equation were studied [1] and then implemented to increase the accuracy of the prediction. In totality 23 features, inclusive of the class label, were considered.

Generic features The generic features were number of variables in the instance, number of variables in the instance, ratio of number of clauses in an instance to the number of variables in an instance, and class label according to whether the instance is satisfiable or not.

Variable–clause bipartite graph A set of features were related to the variable-clause bipartite graph. In this graph variables and clauses in an instance are visualized as nodes and an edge occurs between a variable and a clause each time a variable occurs in that clause. These features are mean of the degree of clause nodes, variation coefficient of degree of clause nodes, minimum of degree of clause nodes, maximum of degree of clause nodes, entropy of degree of clause nodes, mean of degree of variable nodes, variation coefficient of degree of variable nodes, minimum of degree of variable nodes, maximum of degree of variable nodes, and entropy of degree of variable nodes.

Variable–variable graph The next set of features were related to the variable-variable graph, wherein all variables in an instance are visualized as nodes and an edge occurs between variables if they occur together in a clause of the instance at least once. These features are mean of degree of variable nodes, variation coefficient of degree of variable nodes, minimum of degree of variable nodes, and maximum of degree of variable nodes.

Ratio of number of positive to negative literals The next set of features relates to the ratio of number of positive literal to the number of negative literals for every variable in an instance. These features are mean of ratio, variation coefficient of ratio, minimum ratio, maximum ratio, and entropy of ratio.

4.2 Dataset

Feature extraction will be done from the uniform random 3-SAT dataset from the SATLIB repository. This dataset includes labelled files of instances with different number of variables and clauses in the .cnf format. Altogether this constitutes to a total of 6400 instances as described in Table 1.

Table 1 Uniform random 3-SAT dataset constituents

Dataset	Number of instances
uf20-91	1000 instances, all satisfiable
uf50-218/uuf50-218	1000 instances, all sat/unsat
uf75-325/uuf75-325	100 instances, all sat/unsat
uf100-430/uuf100-430	1000 instance, all sat/unsat
uf125-538/uuf125-538	100 instances, all sat/unsat
uf150-645/uuf150-645	100 instances, all sat/unsat
uf175-753/uuf175-753	100 instances, all sat/unsat
uf200-860/uuf200-860	100 instances, all sat/unsat
uf225-960/uuf225-960	100 instances, all sat/unsat
uf250-1065/uuf250-1065	100 instances, all sat/unsat

In uniform random 3-SAT, uniformity refers to the fact that the variables in each of the instances are distributed evenly in all of the instances of a dataset, for the same number of clauses and variables. This results in the mean of variables participating in clauses to be the same for each instance. So the mean of degree of variable nodes in the variable-clause bipartite graph comes out to be the same of all instances in the satisfiable as well as unsatisfiable collective. This also implies that this feature then becomes useless for the classification task. Due to the nature of the dataset chosen, a few other features can also be dropped without the performance of the models for prediction being affected.

5 Training and Testing

There are a variety of machine learning algorithms that are available for the purpose of classification. Those that are picked for this system are the linear classifier, dense neural network classifier, decision tree, random forest classifier, Gaussian naive Bayes, Bernoulli naive Bayes and multinomial naive Bayes. The linear classifier is the most basic machine learning classification model. It uses sigmoid function as its cost function. The linear classifier uses a linear combination of features to differentiate the classes. This restriction makes it difficult for it to form an appropriate decision boundary. Hence, to implement a model that makes use of polynomial features we have selected the dense neural network classifier. The activation function which is used in this model is ReLu. Another non-linear classifier that we used is the decision tree. The decision tree has a problem of high variance, overfitting, because of which it doesn't perform well on instances it has never seen before during training. This drawback is overcome by using the random forest model. The random forest outputs a class based on the value of several decision trees, with considering random subsets

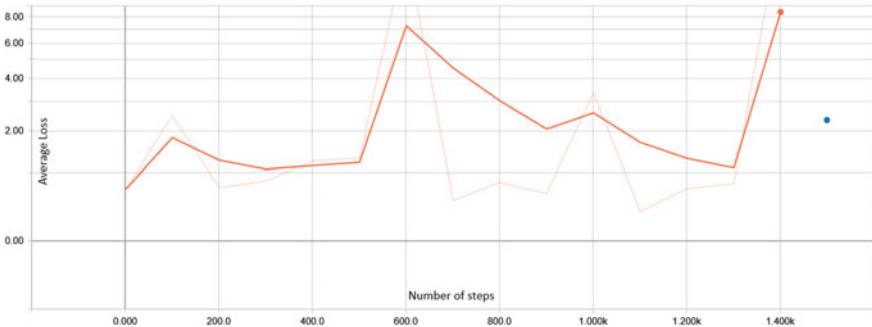


Fig. 1 Average loss for the linear classifier as number of steps of training increase

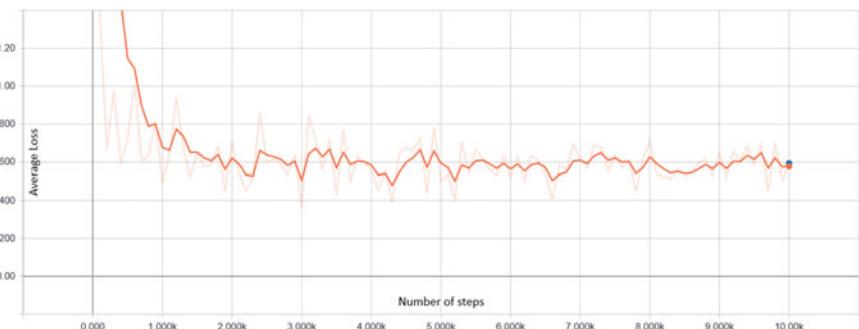


Fig. 2 Average loss for the dense neural network as number of steps of training increase

of the feature set for each tree, which prevents it from overfitting. The last model which we used is the naive Bayes classifier.

The naive Bayes classifier is based on Bayes theorem of probability and assumes that all the features are independent of each other. We have implemented this model just for comparison. The SATZilla features which used are dependent on each other, the naive Bayes model, therefore, is not expected to perform efficiently.

Figures 1 and 2 show the tensorboard visualization of the training set for the given dataset. The models which were visualized are the linear classifier and the dense neural network. The linear classifier is trained for 1500 steps starting with an average loss that lies between 0.00 and 2.00. As the training process progresses the average loss goes on increasing reaching the global maximum at the 1500 steps. On the other hand, the dense neural network's maximum loss occurs when the training process begins and goes on decreasing as the process is advanced reaching a global minimum at 10,000 steps.

6 Implementation

The cnf files from the datasets were sieved and put into a Python list. Each list containing a list of instances, which included a list of clauses within each instance. A clause was described by its three constituent literals as integers. A positive integer corresponding to a variable and a negative corresponding to its conjugate. These lists were then used to extract features to be stored in the feature dictionary. Each set of features was calculated as discussed before.

These feature dictionaries were concatenated and put into pandas dataframes to be modelled using the chosen machine learning models. A 0.33 split was made to the dataset to produce the test and training sets using scikit-learn. The linear classifier and dense neural network classifier was implemented using TensorFlow. The decision tree, random forest with 30 estimators and naive Bayes models; Gaussian, Bernoulli and multinomial were implemented using scikit-learn. The training of the linear classifier and dense neural network were visualized using TensorBoard. The f1-scores for the satisfiable, unsatisfiable classes and their weighted average from the testing were recorded. The scores were plotted using matplotlib.pyplot to produce bar graphs and line graphs.

7 Quantifying Results

To quantify the results obtained f1-score was used. Accuracy, when used, isn't entirely reflective of the performance of the algorithms. Whereas, precision and recall, whose harmonic mean give f1-score are. This is especially relevant to the data to be modelled in this system, since the flip in even a single bit can render into changing the class that an equation belongs to. If a test set has, say, 100 instances off of which 85 are satisfiable and if the trained model predicts the entire set to be satisfiable, it'll still produce an accuracy of 0.85 which represents a decent performance from the model but this, in actuality, isn't the case.

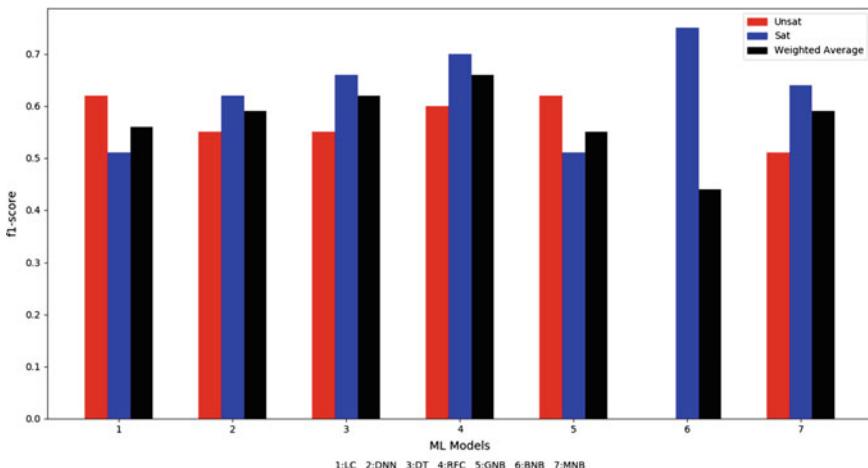
Precision, recall and f1-score are given for each class label and then a weighted average, according to support, of each of these class values is found. Considering the confusion matrix, precision and recall are defined in terms of true positives, false positives, true negatives and false negatives. Precision is the ratio of true positives to the sum of true positives and true negatives for a class. Recall is the ratio of true positives to the sum of true positives and false negatives of a class.

8 Discussion and Inference

The f1-scores show that the random tree classifier performs the best on both the classes, SAT and UNSAT, from amongst the models employed. This is followed by the decision tree and the dense neural network classifier. The values of weighted

Table 2 Models chosen with corresponding f1-scores

Model	Weighted average of f1-scores
Linear classifier	0.56
Dense neural network	0.59
Decision tree	0.62
Random forest	0.66
Gaussian naive Bayes	0.55
Bernoulli naive Bayes	0.44
Multinomial naive Bayes	0.59

**Fig. 3** The f1-scores for each model with respect to the SAT class, UNSAT class, and weighted average of these f1-scores

average of f1-scores are enlisted in Table 2 and visualised in Fig. 3 with an addition of f1-scores for each class.

In Fig. 3, LC stands for Linear Classifier, DNN for dense neural network, DT for decision tree, RFC for random forest classifier, GNB for gaussian naive Bayes, BNB for Bernoulli naive Bayes, MNB for multinomial naive bayes.

Figure 4 shows the performance of each of the models as a line graph as the number of variables in the instances of the test set increases. This performance is expressed as a ratio of correctly predicted instances to the total number of test set instances for each every variable. All the models perform well on the 20 variables, 91 clauses dataset since it contained only satisfiable instances. During training, only satisfiable instances were which created an inherent bias and had the models predict the satisfiable class for all test instances. The random forest classifier gives an accuracy of above 0.50 throughout as the number of variables increase.

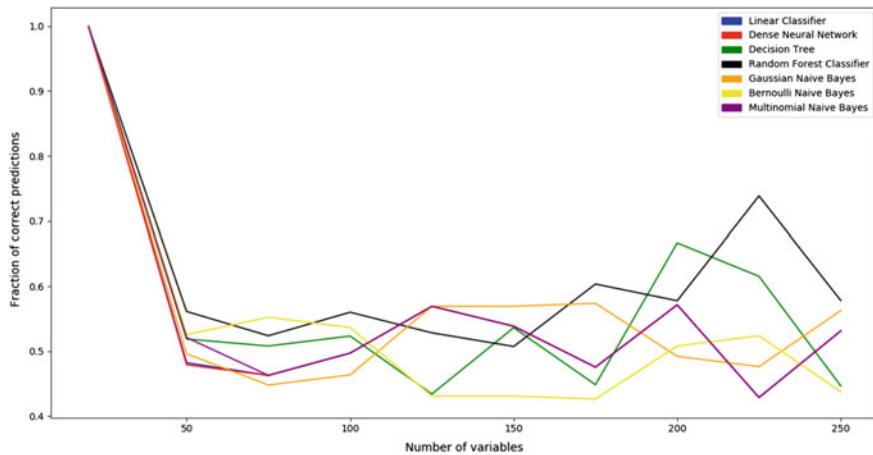


Fig. 4 Fraction of predictions made correctly by each model as number of variables and clauses in the instances increase

The naive Bayes models need an independent set of features that equally contribute to the classification task. Both these conditions were not true in case of the features chosen for this system. Having said that, multinomial naive Bayes performed better than the other two naive Bayes models.

The models performed does not perform better and this could be owing to several factors, including the choice of feature set, the choice of the feature set with respect to the dataset chosen, the choice of hyperparameters for the models or the choice of models itself.

9 Future Scope

In the future, we would want to employ more machine learning models along with other techniques (Graph Neural Network [2], randomized algorithm approach) to further our work in finding the best algorithm for solving the 3-CNF-SAT problem. It must be noted that there are many different variations of a 3-CNF-SAT equation and each can give optimum results with different algorithms and hyperparameters.

Further we would work to expand the datasets used and inherently the set of features used in relevant models. This could mean that we would have to come up with our own datasets.

Comparing models on the basis of parameters of time and space complexities in addition to f1-scores of prediction would help us better evaluate and compare the models used. Additionally, comparing these machine learning models with the traditionally used approximate algorithms would provide us with a better comparison of performance of the various machine learning models.

Finally, just being able to determine if equations are satisfiable or not provides us with a lot of valuable information. Being able to find a satisfiable assignment for these satisfiable equations is also an NP-complete problem in itself. Being able to solve this problem of finding an assignment provides us with a definitive solution to a satisfiability problem.

References

1. D. Devlin, B. O'Sullivan, Satisfiability as a Classification Problem Grant No. 05/IN/I886 Cork Constraint Computation Centre Department of Computer Science, University College Cork, Ireland Supported by Science Foundation Ireland. <http://www.cs.ucc.ie/~osullb/pubs/classification.pdf>
2. D. Selsam, M. Lamm, B. Bunz, P. Liang, D.L. Dill, Learning a SAT Solver from Single-Bit Supervision Submitted on 11 Feb 2018, last revised 5 Jan 2019. Department of Computer Science Stanford University Stanford, CA 94305 and Microsoft Research Redmond, WA 98052. <https://arxiv.org/pdf/1802.03685.pdf>

A Design and an Implementation of Forecast Sentence Extractor



Benyatip Srichareon, Suparerk Manitpornsut and Prapas Pongdamrong

Abstract Strategic planning is a practical approach for researchers to conduct the STEEP analysis. One of the most promising approaches for strategic planning is the Foresight Framework. In the very first steps of Foresight Framework, however, the environmental scanning is involved. This process is time-consuming since a very large amount of data must be explored. To alleviate the time in such a process, this study proposes a design and an implementation of the forecast sentence extractor by using natural language processing and machine learning algorithm. The proposed algorithm digests a long article and then provides a short list of forecast sentences. Three feature selection approaches are tested. From the experimental studies, the accuracy of the proposed algorithm is up to 85.10%.

Keywords Foresight Framework · Machine learning · Classification · Natural language processing

1 Introduction

Business data analytics has been widely adopted in enterprises, serving as a tool for production planning, sale monitoring, marketing campaign monitoring, employee performance measuring, customer loyalty program planning, etc.

Traditional business data analytics is mostly based on the “internal data”—data within the organization, e.g., data in spreadsheets and databases. Enterprise tools,

B. Srichareon · S. Manitpornsut (✉)

Department of Computer Engineering and Artificial Intelligence, University of the Thai Chamber of Commerce, Bangkok 10400, Thailand

e-mail: suparerk_man@utcc.ac.th

B. Srichareon

e-mail: benyatip_lua@utcc.ac.th

P. Pongdamrong

AI Lab, Betimes Solutions, Bangkok 10260, Thailand

e-mail: prapas.p@betimes.biz

such as customer relationship management system (CRM), enterprise resource planning (ERP), and business intelligence system (BI), can exceptionally cope with this kind of data.

Nonetheless, online social networks, e.g., Facebook, Twitter, and Instagram, have been extensively embraced for both personal usage and enterprise communication. Many organizations welcome this new approach of communication between themselves and their customers. As a result, social listening becomes a great tool for enterprise to listen to the voices of customers. This process includes the interpretation of customers' questions and comments, the trends of number of "Shares" and "Likes", the age range of fan page visitors, etc. These data can be considered as "external" data—data generated by others. While the majority of internal data are structured, most of external data are unstructured—no predefined data model or structure. Natural language processing (NLP), therefore, is usually exercised to extract specific meaning from such data, e.g., sentiment and entities.

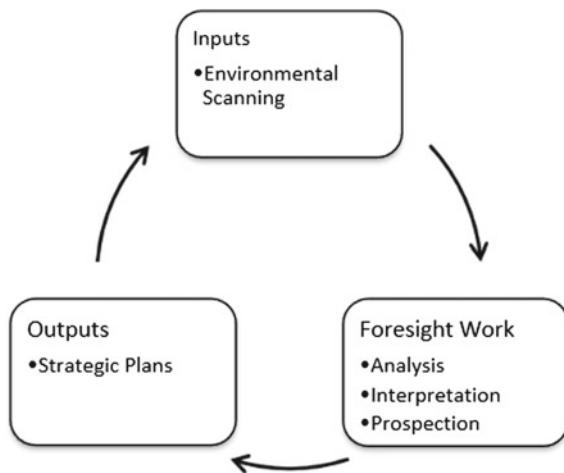
By continuously using both internal and external data in business data analytics, enterprise can answer the questions, e.g., "What happened in the last quarter?" or "What is happening today?"

Based on merely historical data, however, the aforementioned approach failed to answer the "long term" question, e.g., "What will happen in the next five years?"

To answer such kind of questions, the strategic planning methodology could be utilized. Foresight Framework [1] is one among the potential approaches for strategic planning. As shown in Fig. 1, there are three major phases in Foresight Framework: Inputs, Foresight Work, and Outputs.

Conventional approach in input phase is Delphi method [2], which is the polling of experts' opinions over several rounds to generate consensus about particular topics or issues.

Fig. 1 Steps in Foresight Framework



There are three steps in the Foresight Work phase: analysis, interpretation, and prospection. Data from the input phase are analyzed by using clustering analysis, pattern analysis, and trend analysis.

During the interpretation process, trends are used in STEEP Analysis (Social, Technological, Economic, Environmental and Political Analysis) to delineate *driving forces* of events of interest, for example, culture differences, alpha generation, big data, global recession, global warming, and political intervention [3].

After interpreting the driving forces, the scenarios are developed in the prospection process. Scenarios can be categorized into possible, plausible, probable, and preferable scenarios [4].

Finally, in the output phase, strategic plans are generated to accommodate the scenarios. To adjust the plan, all these three phases may be repeated. Note that, traditionally, all phases are done manually by strategic planners.

Presently, Delphi method is almost inapplicable due to too few experts in the same vicinity, high operation cost (e.g., experts traveling cost, organizing cost), its time-consumed characteristic, etc. We can overcome these problems to a certain extent by automatically gathering related articles from experts over the world through the Internet by using a web crawler. However, a new problem arises, too many articles to be explored by strategic planners. To alleviate this problem, we propose the Forecast Sentence Extractor, a tool that can extract “forecast” sentences from a long article. Forecast sentences can help strategic planners to decide if such an article is related to their issue of interest. The examples of forecast sentences are as follows:

The shining beacon in this despotic dashboard situation is Renault-Nissan-Mitsubishi's recent announcement that it will hand over its center stack infotainment system to Android starting in 2021. [5]

The global IoT in healthcare market size is projected to reach USD 534.3 billion by 2025 expanding at a CAGR 19.9% over the forecast period, according to a new report by Grand View Research, Inc. [6]

In the next section, the related works are discussed, i.e., natural language processing and machine learning APIs. The proposed system is explained in Sect. 3. The experimental studies and discussion are given in Sect. 4. Finally, the conclusion and future works are presented in Sect. 5.

2 Related Works

2.1 Natural Language Processing

Fundamental operations of natural language processing (NLP) include tokenization, part of speech tagging, name entity recognition, semantic analysis, and sentiment analysis.

In recent years, NLP has gained great interest due to social networks and demands in text processing. There are many NLP tools in both proprietary and open-source APIs, e.g., Stanford NLP [7], NLTK [8], Microsoft LUIS [9], and IntelleXer [10].

Since the number of articles in input phase of Foresight Framework is extremely high, manually classifying these articles by strategic planners takes time and is inefficient. NLP together with machine learning algorithm can enhance this process by classifying sentences in the articles if they are forecast sentences. By looking at the forecast sentences, strategic planners can decide if the articles are of their interests or if they should peruse the full articles.

Due to its open-source license, reproducible research results, and well-written examples, Stanford NLP is utilized in the implementation of the proposed system. More details of Stanford NLP APIs are explained in the later section.

2.2 *Machine Learning APIs*

Currently, machine learning (ML) is one of the very hot topics in computer science. It has been integrated into a lot of applications, for example, keyword extraction by using text clustering technique, user comment classification by using text classification algorithm, and item recommendation in online store.

There are several machine learning APIs available for academic usage and for commercial integration, e.g., Weka [11], RapidMiner [12], Azure Machine Learning Services [13], Google Cloud AutoML [14], Apache Mahout [15], and Apache Spark [16].

In this research paper, Apache Mahout is integrated into the proposed system. Not only does it offers standard implementation of machine learning algorithms, it is also very flexible to run either as the standalone application or as the distributed application on Hadoop cluster [17]. In the next section, detail of the implementation is explained.

3 Proposed System

Since the complete system of Foresight Framework is very complicated, the proposed system is only one part in input phase with the following assumptions:

- Articles are already collected from trusted sources and only their contents are saved as text format in the file system. Other tags, e.g., HTML tags are removed. These articles are the inputs of the proposed system.
- The proposed system extracts each article into sentences and then classifies them if they are forecast sentences. The output from the proposed system, therefore, is a list of forecast sentences.

As discussed previously, Stanford NLP and Apache Mahout are used as the NLP and ML APIs. However, the abstract factory design pattern is applied to loosely couple ForecastSentenceExtractor from the underlined NLP and ML APIs. As shown in Fig. 2, ExtractorFactory, CoreNLP and CoreML are interfaces, while StanfordCoreNLP and Mahout-CoreML are our wrapper classes for Stanford NLP and Apache Mahout APIs, respectively. The concrete implementation of ExtractorFactory is SMEx-tractorFactory. With this design, NLP engine and ML APIs can be effortlessly changed in the future.

There are two key methods in ForecastSentenceExtractor class: getForecastSentences and isForecastSentence. The pseudocode of getForecastSentences is given in Fig. 3.

Article is analyzed and parsed into sentences by parser and an object from CoreNLP. Then, each sentence is tested against isForecastSentence function. Strategy design pattern (not included in Fig. 2) is implemented for isForecastSentence function to support three different feature selections and allow different learningAlgorithm to be implemented.

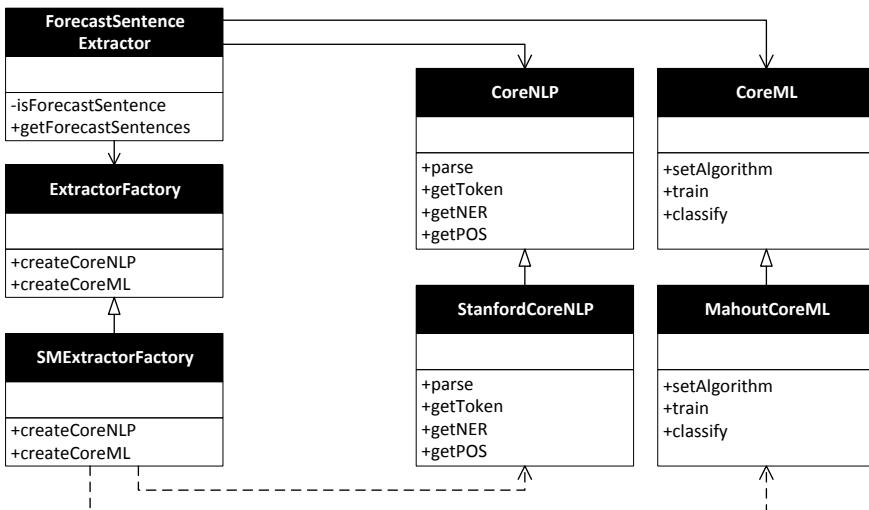


Fig. 2 Class diagram of the ForecastSentenceExtractor

```

Function getForecastSentences(A : Article)
  Sentence [] sentences = parser.parse(A)
  Foreach(s : Sentence in sentences)
    If s.isForecastSentence(), then
      candidateList.add(s)
    EndIf
  End
  Return candidateList
  
```

Fig. 3 Pseudocode for getForecastSentences

PoS Only is the first feature selection approach of `isForecastSentence`, listed as pseudocode in Fig. 4. Only word with the following PoS tags are injected into data vector v: MD, VB, VBD, VBG, VBP, and VBZ. These tags are the labels for various forms of verbs and modals. For the full list of PoS tags, please visit the Penn Treebank Project [18]. Data vector v is the random-access sparse vector and `learningAlgorithm` is the classification algorithm from CoreML that can classify data vector v into categories according to the training data model. In this case, there are only two categories: *Forecast* and *Nonforecast*.

During the training data preparation, we notice that many of forecast sentences express expected year. We, therefore, add the year (if any) into data vector v as shown in Fig. 5. By using named entity recognition (`getNER`) in CoreNLP, year can be extracted from the sentence and checked if it is the future, not in the past. This feature selection is named as *PoS and Year*.

The last approach of feature selection is *Keyword and Year*. In Fig. 6, all “keywords” in a sentence are taken into account as the context to build the data vector v. In addition, `year` (if any) is treated as the bias. The keywords are any words not in the Stop Word List. Sample words in Stop Word List are as follows:

```
Function isForecastSentence (s : Sentence)
    Token [] tokens ← getToken()
    Foreach(t : Token in tokens)
        pos ← getPoS()
        If isVerb(pos), then
            encoder.addToVector(t.Text, v)
        EndIf
    End
    p = learningAlgorithm.classify(v)

    Return p
End
```

Fig. 4 `isForecastSentence`: PoS Only

```
Function isForecastSentence (s : Sentence)
    Token [] tokens ← getToken()
    Foreach(t : Token in tokens)
        pos ← getPoS()
        year ← getNER()
        If isVerb(pos) or isFuture(year), then
            encoder.addToVector(t.Text, v)
        EndIf
    End
    p = learningAlgorithm.classify(v)

    Return p
End
```

Fig. 5 `isForecastSentence`: PoS and Year

```

Function isForecastSentence (s : Sentence)
    Token [] tokens ← getToken()
    Foreach(t : Token in tokens)
        If (isKeyword(t)), then
            encoder.addToVector(t.Text, v)
        EndIf
        pos ← getPoS()
        year ← getNER()
        If isVerb(pos) or isFuture(year), then
            bias.addToVector(t.Text, v)
        EndIf

    End
    p = learningAlgorithm.classify(v)

    Return p
End

```

Fig. 6 isForecastSentence: Keyword and Year

- *Article*: a, an, the.
- *Preposition*: e.g., in, on, under.
- *Pronoun*: e.g., he, she, they.
- *Some Words*: e.g., only, too, very, how, why, both, some, etc.
- *Special Characters*: e.g., full stop, tab, space, exclamation mark, question mark, etc.

4 Experimental Studies

4.1 Classification Parameters

Online logistic regression algorithm from Apache Mahout APIs is assigned to the learningAlgorithm object. As discussed previously, strategy design pattern is utilized to allow other algorithms to be used in the future. In our experimental studies, parameters are set as shown in Table 1.

Table 1 Parameters in logistics regression

Parameter	Value
Alpha	1.0
Decay exponent	0.9
Step offset	10,000
Lambda	3.0×10^{-5}
Learning rate	20

Table 2 Features selection

Feature selection	Accuracy
PoS only	73.72%
PoS and Year	79.44%
Keyword and Year	85.10%

4.2 Training and Testing Data

A thousand of *Forecast* and *Nonforecast* sentences (500 each) are collected from various trusted sources, e.g., Harvard Business Review [19], Shaping Tomorrow [20], World Economic Forum [21], Thai Productivity Institute [22], etc. Fivefold cross validation technique is applied, thus eighty percent of all those sentences being used as the training data and the rest as testing data.

4.3 Experimental Results

As explained in the previous section, there are three feature selection approaches PoS Only, PoS and Year, and Keyword and Year.

In Table 2, the experimental results among these feature selection approaches are compared. Comparing between PoS Only and PoS and Year, the accuracy of the classification can be improved approximately 6% when Year as the bias is applied. Moreover, in Keyword and Year, the accuracy can be further improved to 85.10% when the Keyword is applied to build data vector v as well as the PoS and Year are used as the bias. A recall for forecast sentences is also tested. The results are 86%, 87%, and 97% from PoS Only, PoS and Year, and Keyword and Year, respectively.

5 Conclusion and Future Works

The forecast sentences extraction, as a part of environmental scanning of the Foresight Framework, is proposed. The logistic regression with three feature selection approaches is applied as the machine learning algorithm. With the Keyword and Year feature and logistic regression algorithm, 85.10% accuracy can be achieved.

However, in the advent of artificial intelligence age, there are plenty of classification algorithms, e.g., decision tree, Bayesian and its derivatives, random forest, etc. Additionally, there are various APIs available in the market. Among those existing APIs, Apache Spark, and TensorFlow are the most promising candidates. We, therefore, will explore these two engines in our future works.

Furthermore, this research is only a part of the input phase in Foresight Framework. More advanced mechanisms are required to fulfill the requirement of Foresight Work and outputs phases. That is the direction of our future research.

Acknowledgements The authors aspire to thank Mr. Tossapol Ramingwong and his team from Thailand Productivity Institute (FTPI), Thailand, for providing Foresight Framework concept and requirement as well as Dr. Sumavasee Salasuk and her team from Digital Economy Promotion Agency (DEPA), Thailand, for delineating Foresight Framework requirement in DEPA point of view. Special thanks go to Mr. Suchart Imbunchon, Mr. Kampon Hannaruechai, Mr. Mahaysak Kanignant, Mr. Khamtipong Damkham, and development team at AI Lab, Betimes Solution Co., Ltd. for helping in requirement analysis, data acquisition, and cleansing processes, as well as project funding. Last, but not least, the authors want to thank all faculty members in the Department of Computer Engineering and Artificial Intelligence, UTCC, for providing challenging comments and creative discussion.

References

1. J. Voros, A generic foresight process framework. *Foresight* **5**(3), 10–21 (2003)
2. O. Helmer-Hirschberg, Analysis of the Future: The Delphi Method, CA (1967), <https://www.rand.org/pubs/papers/P3558.html>. Accessed 01 Nov 2018
3. Managing Business at Speed of Change. Thai Productivity Institute (2018), https://www.ftpi.or.th/download/seminar-file/Speed-of-Change_E0B8AAE0B896E0B8B2E0B89AE0B8B1E0B899E0B980E0B89EE0B8B4E0B988E0B8A1.pdf. Accessed 22 Apr 2018
4. J. Voros, Foresight Primer—Thinking Futures, Thinking Futures (2001), <https://thinkingfutures.net/foresight-primer>. Accessed 17 May 2018
5. B. Cooley, Why can't my passenger enter car GPS navigation? Cars increasingly want us to use voice while driving, Roadshow by CNET (2018), <https://www.cnet.com/roadshow/news/why-cant-my-passenger-enter-car-gps-navigation/>. Accessed 01 Apr 2019
6. IoT in Healthcare Market Worth \$534.3 Billion By 2025|CAGR: 19.9%, Grandviewresearch.com (2019), <https://www.grandviewresearch.com/press-release/global-iot-in-healthcare-market>. Accessed 01 Apr 2019
7. The Stanford Natural Language Processing Group, <https://nlp.stanford.edu/>. Accessed 12 May 2018
8. Natural Language Toolkit, <http://www.nltk.org/>. Accessed 12 May 2018
9. LUIS (Language Understanding)—Cognitive Services—Microsoft Azure, luis.ai, <https://www.luis.ai/home>. Accessed 12 May 2018
10. IntelleXer—Text Mining Solutions for Everyone, <http://www.intelleXer.com/>. Accessed 12 May 2018
11. E. Frank, M.A. Hall, I.H. Witten (2016). *The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques*, 4th edn. (Morgan Kaufmann, 2016)
12. I. Mierswa, R. Klinkenberg, RapidMiner Studio (2018), <https://rapidminer.com/>. Accessed 12 May 2018
13. Microsoft Azure Cognitive Services (2018), <https://azure.microsoft.com/en-us/services/cognitive-services/>. Accessed 12 May 2018
14. Google Cloud AutoML Natural Language|AutoML Natural Language|Google Cloud. Google Cloud (2019), <https://cloud.google.com/natural-language/automl/docs/>. Accessed 10 Mar 2019
15. D. Lyubimov, A. Palumbo, Apache Mahout: Beyond MapReduce (2016), <https://dl.acm.org/citation.cfm?id=3019198>. Accessed 17 May 2018
16. M. Zaharia, R.S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M.J. Franklin, A. Ghodsi, J. Gonzalez, S. Shenker, I. Stoica, Apache Spark: a unified engine for big data processing. *Commun. ACM* **59**(11), 56–65 (2016). <https://doi.org/10.1145/2934664>

17. Apache Hadoop, hadoop.apache.org (2019), <https://hadoop.apache.org/>. Accessed 12 May 2018
18. Penn Treebank P.O.S. Tags, Department of Linguistics, University of Pennsylvania (2019), https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html. [Accessed: 12- May- 2018]
19. Harvard Business Review—Ideas and Advice for Leaders, hbr.org, <https://hbr.org/>. Accessed 20 Feb 2019
20. Shaping Tomorrow, Shaping Tomorrow, <https://shapingtomorrow.com/home>. Accessed 20 Feb 2019
21. World Economic Forum, World Economic Forum (2019), <https://www.weforum.org>. Accessed 20 Feb 2019
22. Foresight, Center of Excellence for Foresight (2019), <http://fis.ftpi.or.th/>. Accessed 20 Feb 2019

Low Complexity Antenna Selection Scheme for Spatially Correlated Multiple Antenna Cognitive Radios



Sonali Chouhan and Tinamoni Taye

Abstract In the spectrum sharing cognitive radio networks, spectrum sensing using multi-antenna can improve sensing performance by exploiting spatial diversity. But because of multipath fading, the spatial correlation arises which depends on the antenna spacing and angle of arrival. This spatial correlation degrades the sensing performance significantly. In this paper, a low complexity antenna selection scheme has been proposed to reduce the effect of spatial correlation and enhance the system sensing performance. In the proposed scheme, all possible less correlated antenna combinations are identified. The total sensing sub-slots are divided by the antenna combinations considered for sensing. In one sensing sub-slot one combination is being used. The theoretical results are verified with the simulations. The simulation results show that the proposed scheme outperforms conventional antenna selection schemes and detection probability approaches unity.

Keywords Cognitive radio · Antenna selection · Spectrum sensing · Spatial correlation

1 Introduction

Increasing demands for radio spectrum and continuous additions of new standards in wireless communication introduces spectrum scarcity problems. To solve this problem Cognitive radio (CR) technology emerges as a promising technology which allows the coexistence of licensed users and unlicensed users by sharing the available spectrum in an opportunistic manner. It promises an efficient utilization of the available radio spectrum. In this coexistence, priority is given to licensed users. To guarantee the priority to the licensed user (primary users (PUs)) and to avoid inter-

S. Chouhan (✉) · T. Taye
Indian Institute of Technology, Guwahati 781039, Assam, India
e-mail: sonali@iitg.ac.in

T. Taye
e-mail: tayetina50@gmail.com

ference, unlicensed users (secondary users (SUs)) must be able to sense the spectrum hole efficiently as well as maintain their transmission reliably [1].

In this scenario, the most important function of the CR is spectrum sensing. In practice, the presence of fading and shadowing adversely affect the sensing performance [2]. To improve the sensing performance, multiple antennas are used to capitalize on the spatial diversity [3]. In the multi-antenna scheme, all the antennas are used for simultaneous sensing and it improves the system performance, but spatial correlation due to antenna spacing degrades the system performance of the energy detector [1]. Spectrum sensing for multi-antenna cognitive radio using the generalized likelihood ratio test [4] outperforms the standard energy detector, but it is too complex to implement. Antenna selection can be implemented either at the SU transmitter or at the receiver. The authors in [5] implemented the antenna selection at the transmitter and proposed to select one out of N antennas. Antenna correlation was not taken into account. The multiple transmit antenna selection is considered in [6, 7]. For optimal antenna selection, a full-scale search was conducted over all the possible combinations in [6]. This increases implementation complexity. The work does not consider the spatial correlation among the selected antennas. To accomplish the task of spectrum sensing, antenna selection at the receiver is needed. To improve the system performance of a multi-antenna system, receiver antenna selection based spectrum sensing is proposed in [8]. The author proposed an antenna selection scheme where certain number of selected set of antennas are being used for sensing in a fixed time period. It shows some improvement, but it does not exploit all possible antenna combinations to reduce the spatial correlation effect. A single receiver antenna is selected in [9, 10].

In this paper, we propose a multiple antenna selection scheme which takes into account the effect of spatial correlation and improves the spectrum sensing performance by simultaneous sensing. In such a scheme the RF-chains required at the receiver are less than the receiver antennas, hence reduces the receiver complexity. In the proposed scheme, the maximum possible combinations of less correlated antennas are identified. Time diversity is also considered. The main contributions of this paper are

- The proposed work exploits all the possible RF combinations of antenna based on the distance effect on spatial correlation.
- The proposal exploits the temporal diversity to improve the multi-antenna CR's sensing performance.
- The proposed scheme is a low complexity solution.

The paper organization is as follows. The system model is presented in Sect. 2, performance analysis is discussed in Sect. 3, Simulation results and analysis of the proposed schemes are described in Sects. 4 and 5 concludes the paper.

2 System Model

Consider a CR system in which licensed PU coexists with the unlicensed SU sharing the same radio spectrum. SU tries to access the radio spectrum opportunistically by sensing PU in the interested radio spectrum. The absence or presence of PU can be formulated by using the binary hypothesis test as

$$\mathcal{H}_0 : y_m(n) = u_m(n) \quad (1)$$

and

$$\mathcal{H}_1 : y_m(n) = \sqrt{\gamma} h_m s(n) + u_m(n) \quad (2)$$

respectively, where $y_m(n)$ is the n th ($n = 1, 2, \dots, l$) discrete received-signal sample at SU receiver antenna m , $m = 1, 2, \dots, M$. M is the number of receiver antenna at the SU. $s(n)$ is the transmitted primary signal which is an i.i.d. random process with zero mean and σ_s^2 variance, i.e., $\mathcal{CN}(0, \sigma_s^2)$ and $u_m(n) \sim \mathcal{CN}(0, \sigma_u^2)$ is the n th noise sample in the sensing time. $s(n)$ and $u_m(n)$ are assumed to be independent over time and correlated over distance and $u_m(n)$ is independent of $s(n)$. γ is the expected SNR at each receive antenna.

The channel coefficient $h_m \sim \mathcal{CN}(0, 1)$ is considered to be different for each antenna but they are spatially correlated. Channel vector \mathbf{h} , having entries of h_m , is given as

$$\mathbf{h} = \sqrt{\mathbf{R}_x} \mathbf{u} \quad (3)$$

where $\mathbf{R}_x = E[\mathbf{h}\mathbf{h}^H]$ is $M \times M$ correlation matrix and \mathbf{u} is the noise vector of length M and $\mathcal{CN}(0, 1)$. The correlation matrix is Toeplitz symmetric matrix whose entries R_{ij} are given as

$$R_{ij} = \begin{cases} 1 & \text{if } i = j \\ \rho & \text{if } i \neq j \end{cases} \quad (4)$$

where spatial correlation coefficient ρ is in the range of $0 \leq \rho \leq 1$.

In the proposed antenna selection scheme, M number of antennas are employed at the sensing node and L RF-chains are available for sensing purpose, where $M \geq L$. Out of ${}^M C_L$ possible combinations, adjacent antenna combinations are not considered because of the distance effect on spatial correlation, i.e., except the highest correlated combinations, all other combinations are considered for sensing. Now, instead of choosing one best combination, we use all chosen combinations. This way we reduce the computational complexity of the antenna selection. All chosen combinations C are not being used for simultaneous sensing, but the total sensing time T is divided into the C sensing sub-slots T_s , $T_s = \frac{T}{C}$. Every antenna combination senses the channel in the T_s sub-slot. Antenna used in one sensing sub-slot may be used in another sub-slot. The illustration of the proposed antenna selection scheme is given in Fig. 1.

Antenna 1	Antenna 2	Antenna 3	Antenna 1	Antenna 1	Antenna 2	M=5
Antenna 3	Antenna 4	Antenna 5	Antenna 4	Antenna 5	Antenna 5	L=2
Ts	Ts	Ts	Ts	Ts	Ts	C=6
Antenna 1		Antenna 2		Antenna 1		M=4
Antenna 3		Antenna 4		Antenna 4		L=2
Ts		Ts		Ts		C=3
Antenna 1			Antenna 2			M=4
Antenna 3			Antenna 4			L=2
Ts			Ts			C=2
Antenna 1				Antenna 1		M=2
Antenna 3				Antenna 3		L=2
Ts=T						C=1

Fig. 1 Antenna selection scheme

After completing the sensing of all antennas, sensing information is being gleaned from different time slots by using weight factor g_m and with energy-detector signal detection is being processed. The energy-detector test statistic is

$$T(y) = \sum_{m=1}^C g_{m1}^2 (\sum_{m=1}^L T_m(y)) \quad (5)$$

where g_{m1} is the weight factor and is given by

$$g_{m1} = \frac{1}{\sqrt{C}}.$$

For the m th antenna the energy-detector test statistics is

$$T_m(y) = \frac{1}{N_1} \sum_{n=1}^{N_1} |y_n(n)|^2 \quad (6)$$

where N_1 is the number of samples in one sensing sub-slot and is given by $N_1 = \frac{N}{C}$.

3 Performance Analysis

The proposed scheme's performance is derived in terms of probability of detection P_d and probability of false alarm P_f . For a large N , by using central limit theorem, the PDF of $T(y)$ under both hypotheses is derived. For M antennas, L RF-chains, and C antenna combinations with a chosen threshold (ϵ), the detection probability, and false alarm probability is

$$P_f(\epsilon) = Q\left(\left(\frac{\epsilon - \sigma_u^2}{\sigma_u^2}\right)\sqrt{N_1 C}\right) \quad (7)$$

and

$$P_d(\epsilon) = Q\left(\frac{\left(\left(\frac{\epsilon}{\sigma_u^2} - \frac{1}{C} \sum_{m_1=1}^C (|h_{m_1}|^2 \gamma + 1)\right) \sqrt{N_1 C}\right)}{\sqrt{1 + \frac{1}{C} (|h_{m_1}|^4 \gamma^2 + 2|h_{m_1}|^2 \gamma)}}\right), \quad (8)$$

respectively. The probability of missed detection P_m can be calculated by $P_m = 1 - P_d$. For a given P_f , the detection threshold ϵ can be calculated as

$$\frac{\epsilon}{\sigma_u^2} - 1 = \left(\frac{Q^{-1}(P_f)}{\sqrt{N_1 C}}\right). \quad (9)$$

For a target detection probability (\bar{P}_d) and false alarm probability (\bar{P}_f) we can express the P_f and P_d , respectively, as

$$P_f = Q\left(\sqrt{\left(1 + \frac{1}{C} (|h_{m_1}|^4 \gamma^2 + 2|h_{m_1}|^2 \gamma)\right)} Q^{-1}(\bar{P}_d)\right) \quad (10)$$

and

$$P_d = Q\left(\frac{\left(Q^{-1}(\bar{P}_f) - \sqrt{\frac{N_1}{C}} \sum_{m_1=1}^C (|h_{m_1}|^2 \gamma)\right)}{\sqrt{1 + \frac{1}{C} (|h_{m_1}|^4 \gamma^2 + 2|h_{m_1}|^2 \gamma)}}\right). \quad (11)$$

4 Simulation Results and Discussion

In this section, we evaluate the sensing performance of the proposed scheme with the following parameters. For the presented results, sensing time considered is $T = 1$ ms. CR employs $M = 4$ antennas and $L = 2$ RF-chains for sensing. Out of M antennas, L antenna combinations are being used for simultaneous sensing in one sensing sub-slot. For this case, as per the proposed antenna selection scheme, $C = 3$ sub-sensing slots are there. We consider the fast varying channel and hence the information in one sensing sub-slot is not correlated with another sensing sub-slot. The SNR of the primary signal is set to be -10 dB. The collected sensing information is combined by using (5). The channel coefficients are derived by using the spatial correlation matrix given in (3). For the near most antennas, spatial correlation is considered as $\rho_1 = 0.9$ and for the farthest antennas, it is considered as $\rho_2 = 0.3$. The probability of detection is set at $P_f = 0.1$ and threshold is calculated by (9).

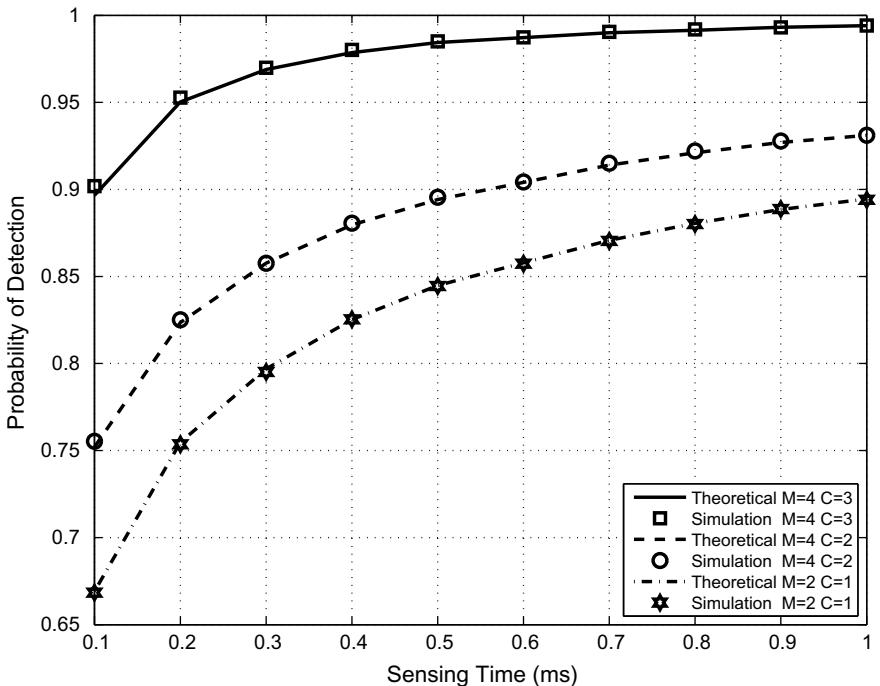


Fig. 2 Sensing performance of CR with $\text{SNR} = -10$ dB, $\rho_1 = 0.9$, $\rho_2 = 0.3$

4.1 Probability of Detection

The plot of probability of detection versus sensing time, shown in Fig. 2, for the proposed antenna selection strategy with $C = 3$ and compare it with the conventional one pair of antenna selection ($C = 1$) and pairs of antenna without repeating antennas in sensing sub-slots ($C = 2$). It is evident that the proposed scheme with $C = 3$ outperforms $C = 2$ and $C = 1$ without increasing receiver complexity. The probability of detection reaches close to unity of the proposed scheme. The Monte Carlo simulation results are very well in accordance with the theoretical results obtained by the analysis presented in the Sect. 3.

In Fig. 3, sensing performance with correlation coefficients $\rho_2 = 0.6$ and $\rho_2 = 0.3$ is shown. It is seen that with less correlation the sensing performance further improves.

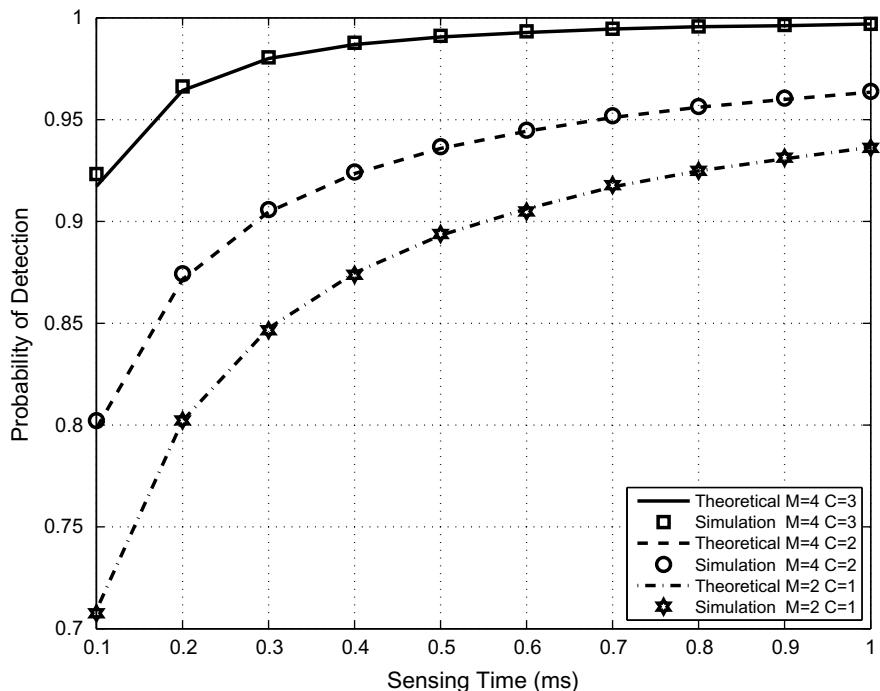


Fig. 3 Sensing performance of CR with SNR = -10 dB, $\rho_1 = 0.6$, $\rho_2 = 0.3$

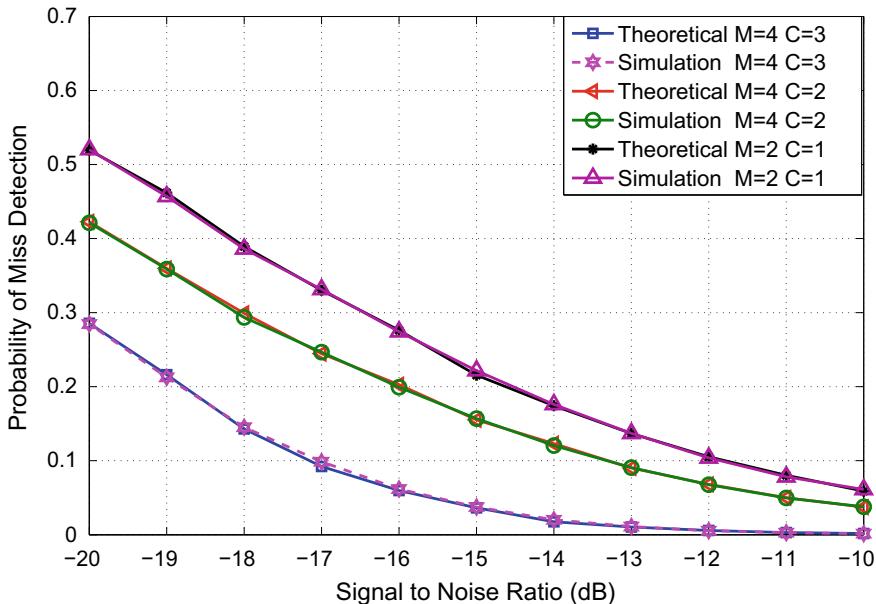


Fig. 4 Sensing performance of CR with respect to SNR

4.2 Effect of SNR on Sensing Performance

The sensing performance in terms of probability of missed detection P_m with respect to various SNR values is shown in the Fig. 4. The proposed solution with $C = 3$ outperforms the other two schemes with $C = 2$ and $C = 1$. It is seen that the performance improves as SNR increases.

5 Conclusions

We proposed an antenna selection scheme to reduce the effect of spatial antenna correlation on the detection performance of secondary users. The proposed scheme exploits temporal diversity as well as spatial diversity and by suitable selection of antennas with the consideration of distance factor on spatial correlation, performance improvement has been shown. It is seen that the proposed scheme with maximum possible antenna combinations outperforms the existing antenna selection schemes. The simulation results are also verified by the theoretical analysis. The performance improvement achieved in this work does not increase the receiver complexity.

References

1. S. Kim, J. Lee, H. Wang, D. Hong, Sensing performance of energy detector with correlated multiple antennas. *IEEE Signal Process. Lett.* **16**(8), 671–674 (2009)
2. A. Ghasemi, E.S. Sousa, Spectrum sensing in cognitive radio networks: requirements challenges and design trade-off. *IEEE Commun. Mag.* **46**, 32–39 (2008)
3. S.M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory* (Prentice-Hall, New Jersey, 1998)
4. R. Zhang, T.J. Lim, Y. Liang, Y. Zeng, Multi-antenna based spectrum sensing for cognitive radios: a GLRT approach. *IEEE Trans. Commun.* **58**(1), 84–88 (2010)
5. K. Tourki, F.A. Khan, K.A. Qaraqe, H. Yang, M. Alouini, Exact performance analysis of mimo cognitive radio systems using transmit antenna selection. *IEEE J. Sel. Areas Commun.* **32**(3), 425–438 (2014)
6. J. Zhou, J. Thompson, I. Krikidis, Multiple antennas selection for linear precoding MISO cognitive radio, in *IEEE Wireless Communications and Networking Conference* (Budapest, 2009), pp. 1–6
7. P. Reba, G.U. Maheswari, S. Babu, Multiple antenna selection for underlay cognitive radio systems with interference constraint. *Wirel. Pers. Commun.* **98**(1), 1505–1520 (2018)
8. S. Wang, Y. Wang, J. Coon, Antenna selection based spectrum sensing for cognitive radio networks, in *International Symposium on Personal, Indore and Mobile Radio Communication* (2011), pp. 364–368
9. S. Narieda, Antenna selection for cyclostationarity detection based spectrum sensing in cognitive radio, in *IEEE Consumer Communications and Networking Conference* (2014), pp. 547–550
10. M. Elsaadany, W. Hamouda, Antenna selection for dual-hop cognitive radio networks: a multiple-relay scenario. *IEEE Trans. Veh. Technol.* **66**(8), 6754–6763 (2017)

Fair Comparative Analysis of Opportunistic Routing Protocols: An Empirical Study



Jay Gandhi and Zunnun Narmawala

Abstract Fair comparative analysis of opportunistic routing protocols plays a vital role in selecting a suitable routing protocol for various applications of opportunistic networks. In this paper, we have analyzed the performance of the routing protocols, namely, EPIDEMIC, Spray and Wait, PROPHET, First Contact, Direct Delivery, MaxProp, WaveRouter, and LifeRouter. The ONE simulator is used for this empirical study. This study measures the performance of protocols based on Delivery Probability, Overhead Ratio, and Average Latency with different mobility models as well as real-world mobility traces. The simulation results surprisingly show that Spray and Wait outperform all the other protocols in almost all scenarios. Further, CAHM mobility model is able to mimic real-world mobility closely resembling real-world mobility traces of different network densities.

Keywords Opportunistic networks · Delay tolerant networks · ONE simulator · Routing protocols · Real-world mobility traces · Synthetic mobility models

1 Introduction

Opportunistic Networks have evolved from Mobile Ad hoc Networks (MANET) in which contemporaneous path between two nodes is not always available [1–3]. Due to this and rapid changes in topologies as a result of mobility of nodes, opportunistic networks use store-carry-forward mechanism. The main objective of routing protocols in these networks is maximizing the message delivery and minimizing the message latency with minimum network overhead [4].

There are mainly two categories of routing protocols: Flooding-based and Forwarding-based [5]. Flooding-based protocols forward a message to “sufficient” number of nodes and hope that the destination node will receive it. Forwarding-based

J. Gandhi · Z. Narmawala (✉)

Computer Science & Engineering, Institute of Technology, Nirma University, Ahmedabad, India
e-mail: zunnun80@gmail.com

J. Gandhi
e-mail: jaygandhi7591@gmail.com

protocols take message forwarding decisions based on the network information collected from other nodes which are coming in contact. In this paper, we study four flooding-based (EPIDEMIC [6], Spray and Wait [7], First Contact [8], and Direct Delivery [8]) and four forwarding-based (PROPHET—Probabilistic Routing Protocol Using History of Encounters and Transitivity [9], MaxProp [10], WaveRouter [3], and LifeRouter [3]) routing protocols by simulating them with mobility models and real-world mobility traces, namely, CAHM (Community Aware Heterogeneous Human Mobility Model) [11], Map-based movement [8], Infocom05 [12], Infocom06 [12], Reality [13], Cambridge [12], and Sassy [14] in ONE simulator [8].

1.1 Opportunistic Routing Protocols

EPIDEMIC routing is flooding-based routing approach [6]. Each node transmits messages to all nodes which come in contact. Spray and Wait algorithm limits this flooding by restricting the maximum number of allowable copies (L) for each message [7]. The PROPHET is probability-based routing algorithm. Each node uses a delivery probability metric to decide whether to forward a message to another node. In the First Contact routing algorithm, a node delivers the message to another node it encounters first and this continues till the message reaches to the destination. In the Direct Delivery approach, the sender node does not forward the message to any other node but delivers the message directly to the destination node if they encounter each other. MaxProp is based on prioritizing the packet transmission schedule and packet drop schedule. Like Prophet, MaxProp also checks if the nodes are likely to meet with the destination node. It uses Dijkstra's algorithm to calculate node to node path using meeting probability. MaxProp removes the delivered message from the network [10]. In WaveRouter, each node receives a message, forwards and stores it for a particular time interval then deletes it. After deleting, the node does not accept that message for a while. All the nodes act in this way to move the message in waves through the grid [3]. LifeRouter is based on the custom Game of Life Simulator [3]. It determines the position of neighbor nodes using radio range. The router uses a parameter, namely, $nmcoun$ t that defines the pair n, m . The message is replicated to the new node or not is based on a value of n and m with number of connected nodes: k [3].

The performance of opportunistic routing protocol can be examined by Delivery Probability, Average Delivery Latency, Overhead Ratio, Buffer Utilization, etc. [8]. These performance results are dependent on various network parameters like Traffic Load, Transmission Range, Node Contact Pattern, Node Density, and many more. Our study on the comparison of the opportunistic routing protocols mainly focuses on Traffic Load (message interval) and Node Contact Pattern. Traffic Load is the amount of traffic generated in the network. If the message interval is less, it means that the traffic load is more. Number of contacts per second is an average number of contacts between nodes in one second. This study also shows how individual protocol performed on different movement models and real-world traces. The study does a

fair comparison of routing protocols to help researchers and network designers in selecting and developing appropriate routing protocol for their opportunistic network scenario.

2 Simulation Setup and Description

2.1 ONE Simulator

For the empirical evaluation, we use the Opportunistic Network Environment (ONE) simulator. It is a java-based tool which has simulation capabilities for the opportunistic network. This framework is capable of generating mobility patterns, simulate real-world traces, routing messages between devices, the graphical user interface to visualize node movement, and reporting of evaluation metrics [8].

2.2 Mobility Models and Real-World Traces

In this paper, we use seven mobility scenarios referred to as CAHM, Map-based Movement, Infocom05, Infocom06, Reality, Cambridge, and Sassy dataset. In which CAHM and Map-based Movement are mobility models whereas the other five are real-world traces. Infocom05, Infocom06, and Cambridge are collected by Haggle project, Reality dataset is from the MIT Reality Mining Project and Sassy is from St Andrews University.

- CAHM: Community Aware Heterogeneous Human Mobility (CAHM) model forms overlapping community structure and moves nodes based on human mobility characteristics derived from real-world mobility traces and human social behavior [11].
- Map-based Movement: In this mobility model, node movement is decided based on a predefined real map. It consists of three movement models, namely, Random Map-based Movement, Routed Map-based Movement, and Shortest Path Map-based Movement [8].
- Infocom05: It is a real-world mobility trace generated from Infocom student workshop. Bluetooth devices were distributed among 50 students who were attending it. Students carried them (imote) in the Infocom 2005 conference for 4 days. The neighborhood scan was done by imote every 2 min [12].
- Infocom06: It is similar to Infocom05 except on a larger scale. The Bluetooth device was carried by 80 students for 5 days in the Infocom 2006 conference. The conference area spanned on three floors, and 34 participants were divided into four groups based on their academic affiliation [12].
- Reality: It is an experiment performed at MIT in which 100 smartphones were given to students and staff for 9 months. These devices were Bluetooth enabled and

- performed device discovery every 5 minutes. This study collected approximately 5,00,000 h of data on the users' communication, location, and devices usage [13].
- Cambridge: In this experiment, two types of contacts were collected, namely, internal and external. Imotes were distributed among 70 students and researchers. Imote recording contact of another imote is known as internal contact and imote recording Bluetooth contact with another external device is known as external contact. The experiment was conducted for 11 days to collect information [12].
 - Sassy: In this experiment, T-mote devices were distributed among 27 staff members of St. Andrews University. The experiment was conducted for 79 days in which staff members carried the device whenever possible. T-mote can detect the device within a range of ~ 10 m. The device encounter events are stored and uploaded to a central database via base stations [14].

2.3 Simulation Parameters

The Table 1 shows simulation parameters used for the simulations done in the ONE simulator.

3 Simulation Results

This section shows the simulation results of opportunistic routing protocols, namely, Direct Delivery, First Contact, EPIDEMIC, PROPHET, Spray and Wait, MaxProp, WaveRouter, and LifeRouter with different mobility models and real-world mobility traces described in Sect. 2. For performance measure, Delivery Probability, Overhead Ratio, and Average Latency metrics are used.

For simulation, we have used Opportunistic Network Environment Simulator (ONE) simulator. The simulation parameters used in the experiments are described in Table 1. Protocol specific parameters for PROPHET and Spray and Wait are as follows: For PROPHET, the values of P_{ini} , β , and γ are 0.75, 0.95, and 0.98, respectively. In Spray and Wait, increasing number of message copies increase the delivery probability but also the overhead. We have kept number of copies at 8, as beyond this value, the protocol's performance improvement is not significant. In LifeRouter, the change in value of n , m plays significant role in performance. In our simulation, the value of n , m is varied between 0 and 5 based on different mobility scenarios.

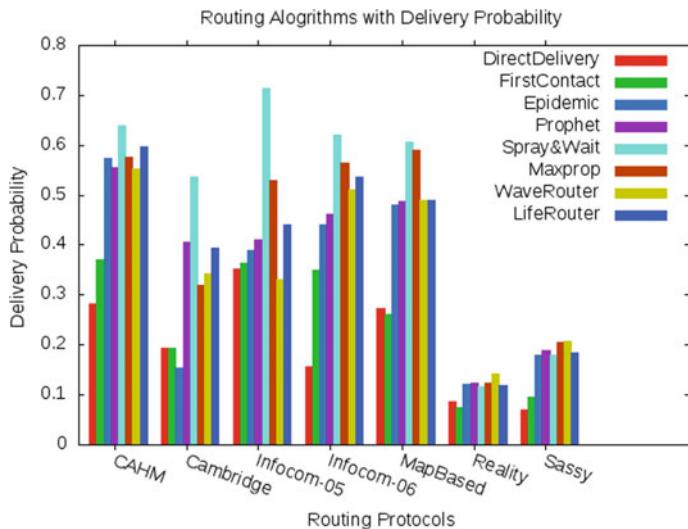
To have fair comparison, all network parameters are kept same for all scenarios. Particular care has been taken in choosing CAHM and Map-based Movement models' parameters to keep average number of node contacts per second with these models is same as Infocom06 traces and is comparable with average number of node contacts per second in Infocom05. For Reality and Sassy, number of contacts per second values are quite low because of low node densities due to larger area. Further, for CAHM and

Table 1 Simulation parameter settings

Simulation time	86,400 s
Interface	Bluetooth
No. of nodes	Mobility Models-100 Infocom05-41, Infocom06-98, Reality-97 Cambridge-54, Sassy-27
Transmit speed	250 KB/s (2Mbps)
Buffer size	100 MB
Message TTL	300 min
Message size	50 KB–1 MB
Message interval	5, 20, 35, 50, 65 s
Warm-up time	1000 s
Avg. number of contacts per second	Reality traces: 0.37 contacts/s All other scenarios: ~1 contact/s

Map-based Movement models, average of 10 simulation runs with different random seeds is taken for all performance measures.

Figure 1 shows that EPIDEMIC, PROPHET, Spray and Wait, and MaxProp have better delivery probability as compared to other single-copy routing protocols. For Reality traces, the delivery probability is low due to very low average number of node contacts. Further, EPIDEMIC performs worst with Reality traces. Due to random flooding, buffers' of nodes which come in contact with each other frequently get full and messages are dropped before they can be delivered to harder-to-reach destination

**Fig. 1** Delivery probability versus routing protocols

nodes. WaveRouter and LifeRouter have also better performance compared to Direct Delivery and First Contact router in terms of delivery probability. MaxProp and Spray and Wait are preferred choice to achieve better delivery probability. Overhead ratio is defined as ratio of total number of messages relayed by all nodes to other nodes which are not destinations and total number of messages successfully delivered to destinations.

As shown in Fig. 2, overhead ratios of EPIDEMIC, WaveRouter, and LifeRouter are quite high in comparison to other protocols as expected. Further, overhead ratio of protocols is quite high in Reality and Sassy traces scenario as compared to other scenarios because of very less number of successfully delivered messages as compared to other scenarios due to very low node density. Also, overhead ratio of PROPHET is quite high as compared to Spray and Wait and MaxProp in all scenarios even though their delivery probabilities are comparable in different scenarios. So, it is evident that Spray and Wait and MaxProp should be the protocols of choice out of all the protocols compared in this paper. The conclusion is contrary to the widely held notion that PROPHET should work better because of its calculation of delivery probability metric.

Average latency is defined as the average time taken by all the messages to reach from their respective sources to their respective destinations. It is evident from Fig. 3 that average latency of all the protocols is comparable in all the scenarios except Reality and Sassy traces scenario because of its lower node density. The Direct Delivery and the First Contact protocol have very high average latency compared to other routing protocols. The simulation result also concludes that the protocols like PROPHET, Spray and Wait, and MaxProp which have better delivery probability also have lower average latency compared to WaveRouter and LifeRouter too.

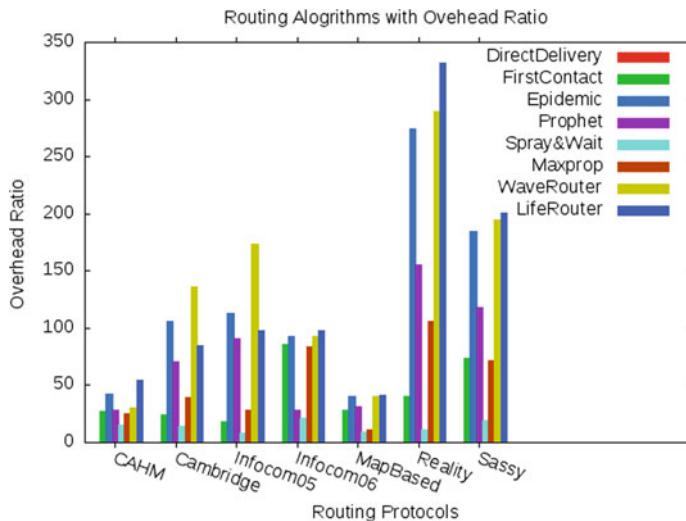


Fig. 2 Overhead ratio versus routing protocols

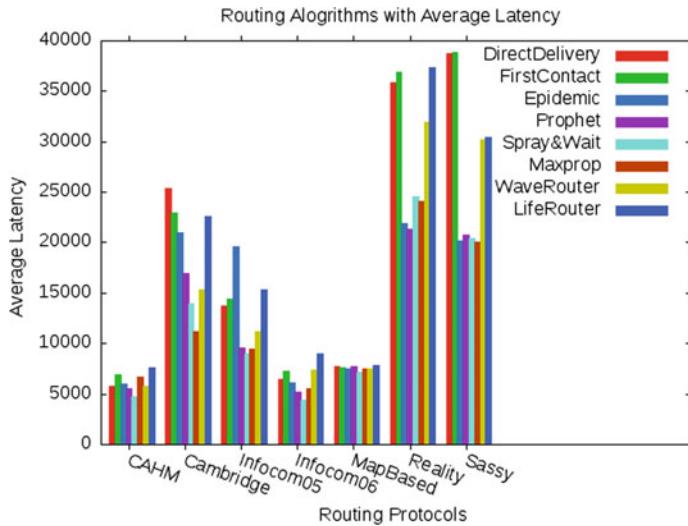


Fig. 3 Average latency versus routing protocols

Figure 4 shows the comparison of routing protocols with different mobility scenarios. The comparison is based on Delivery Probability. From the result, it is clear that CAHM and Map-based Movement models are able to generate mobility patterns closely resembling real-world traces having same average number of node contacts per second.

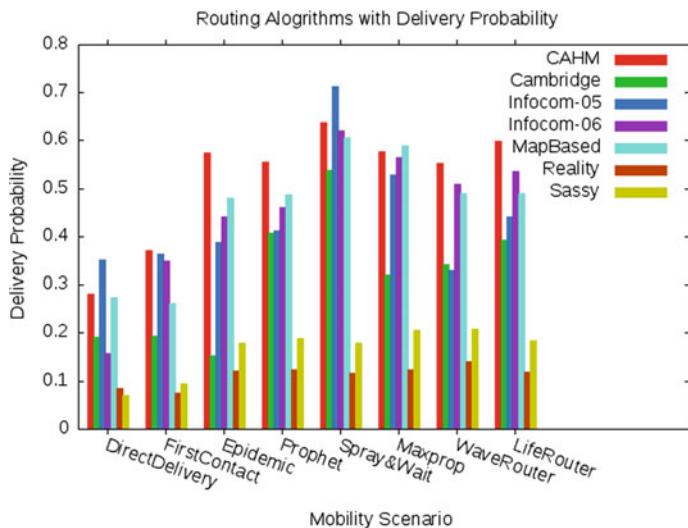


Fig. 4 Delivery probability versus mobility scenarios

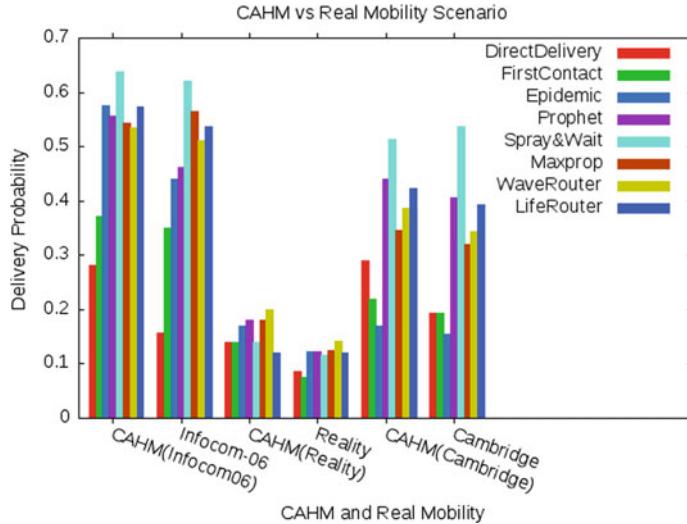


Fig. 5 Delivery probability versus mobility scenarios

For all the above results, average number of node contacts in CAHM was kept same as Infocom06 traces by setting cell size as 82 in CAHM. It is denoted as CAHM (Infocom06) in Fig. 5. To verify that CAHM can generate mobility patterns closely resembling networks of different densities, we have done simulations with average number of node contacts per second same as Reality traces by setting cell size as 160 in CAHM. It is denoted as CAHM (Reality). We have also compared CAHM with one more real trace, namely, Cambridge. It is denoted as CAHM (Cambridge) in Fig. 5. For this comparison, cell size in CAHM is 110. Figure 5 shows the comparison of CAHM with Infocom06, and Reality and Cambridge traces. It is clear from the figure that delivery probabilities of different routing protocols with CAHM (Infocom06) and Infocom06 traces are very similar. Similarly, delivery probabilities of different routing protocols with CAHM (Reality) and Reality traces and CAHM (Cambridge) and Cambridge are also similar. So, it can be concluded that CAHM is able to mimic mobility closely resembling real-world mobility traces of different network densities.

4 Conclusion and Future Directions

In this paper, we studied the performance of opportunistic routing algorithms, namely, First Contact, Direct Delivery, EPIDEMIC, Spray and Wait, PROPHET, MaxProp, WaveRouter, and LifeRouter. The performance of protocols is measured with different mobility models and real-world traces. This study allows us to derive the following conclusions:

- EPIDEMIC performs worst in very low network density.

- WaveRouter and LifeRouter have significant performance improvement than EPI-DEMIC.
- Overhead ratio of all protocols in very low network density is quite high as compared to dense networks.
- Overhead ratio of PROPHET is quite high as compared to Spray and Wait and MaxProp even though their delivery probabilities are comparable. So, Spray and Wait and MaxProp should be preferred over PROPHET.
- CAHM is able to mimic mobility closely resembling real-world mobility traces of different network densities.

In the future, we plan to extend this work with additional mobility models and opportunistic routing protocols with different performance affecting parameters like random seeds, message TTL, message size, and buffer size. We are planning to use machine learning algorithms to understand the mobility patterns of nodes, to identify important nodes in networks like Hub and Gateway and to find probability of transferring messages between two nodes.

References

1. J. Dede, A. Förster, E. Hernández-Orallo, J. Herrera-Tapia, K. Kuladinithi, V. Kuppusamy et al., Simulating opportunistic networks: survey and future directions. *IEEE Commun. Surv. Tutor.* **20**, 1547–1573 (2018)
2. S. Bharamagoudar, S. Saboji, Routing in opportunistic networks: taxonomy, survey, in *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)* (2017), pp. 300–305
3. A. Keränen, E. Hyytiä, J. Ott, M. Desta, T. Kärkkäinen, Evaluating (Geo) content sharing with the one simulator, in *11th ACM International Symposium on Mobility Management and Wireless Access (MobiWac'13)* (2013)
4. R. Cavallari, S. Toumpis, R. Verdone, Analysis of hybrid geographic/delay-tolerant routing protocols for wireless mobile networks, in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications* (2018), pp. 2321–2329
5. V.F.S. Mota, F.D. Cunha, D.F. Macedo, J.M.S. Nogueira, A.A.F. Loureiro, Protocols, mobility models and tools in opportunistic networks: a survey. *Comput. Commun.* **48**, 5–19 (2014)
6. A. Vahdat, D. Becker, *Epidemic Routing for Partially Connected Ad hoc Networks* (2000)
7. T. Spyropoulos, K. Psounis, C.S. Raghavendra, Spray and wait: an efficient routing scheme for intermittently connected mobile networks, in *Proceedings of the 2005 ACM SIGCOMM Workshop on Delay-Tolerant Networking* (2005), pp. 252–259
8. A. Keränen, J. Ott, T. Kärkkäinen, The ONE simulator for DTN protocol evaluation, in *Proceedings of the 2nd International Conference on Simulation Tools and Techniques* (2009), p. 55
9. A. Lindgren, A. Doria, O. Schel, #233, Probabilistic routing in intermittently connected networks. *SIGMOBILE Mob. Comput. Commun. Rev.* **7**, 19–20 (2003)
10. J. Burgess, B. Gallagher, D. Jensen, B.N. Levine, Maxprop: Routing for vehicle-based disruption-tolerant networks, in *25th IEEE International Conference on Computer Communications. Proceedings INFOCOM* (2006), pp. 1–11
11. Z. Narmawala, S. Srivastava, Community aware heterogeneous human mobility (cahm): model and analysis. *Pervasive Mob. Comput.* **21**, 119–132 (2015)
12. S. James, G. Richard, C. Jon, H. Pan, D. Christophe, C. Augustin, CRAWDAD dataset cambridge/haggle (v. 2009-05-29) ed (2009)

13. N. Eagle, A. Pentland, Reality mining: sensing complex social systems. *Pers. Ubiquit. Comput.* **10**, 255–268 (2006)
14. B. Greg, H. Tristan, R. Devan, B. Martin, B. Saleem, CRAWDAD dataset st_andrews/sassy (v. 2011-06-03) ed (2011)

Distributed Optimal Power Allocation Using Game Theory in Underlay Cognitive Radios



Bhukya Venkatesh, Nadella Bala Sai Krishna and Sonali Chouhan

Abstract In underlay cognitive radio networks (CRNs), primary licensed user and secondary unlicensed users use the same spectrum simultaneously by adjusting transmit power of secondary user. Such CRN consists of many secondary base stations (sec-BSSs), primary base stations (prim-BSSs), secondary user terminals (sec-UTs), and primary user terminals (prim-UTs). In this case, the major concern is to limit the interference at each prim-UT. This concern becomes a constraint for the sec-BSSs in assigning transmit powers. In this paper, we develop the complication of power allocation to the sec-BSSs as a concave game where there is no cooperation and communication between sec-BSSs. The sec-BSSs are considered to be players and the interference constraints are imposed by the prim-UTs. Unlike using the traditional Nash Equilibrium for equilibrium selection, we use the Normalized Nash Equilibrium, which is found by solving the necessary KKT conditions and checking the existence of the Lagrangian multipliers. The problem is further improvised by considering the battery leakage. The simulation results demonstrate the optimal power allocation for the sec-BSSs taking the battery leakage into account.

Keywords Cognitive radio · Power allocation · Game theory · Normalized Nash equilibrium · Optimization

1 Introduction

In cognitive radio networks (CRNs), due to spectrum scarcity, the spectrum allocated to licensed user is being shared by the unlicensed user. Licensed users consist of a primary base station (prim-BS) and primary user terminals (prim-UTs), whereas

B. Venkatesh · N. Bala Sai Krishna · S. Chouhan (✉)
Indian Institute of Technology, Guwahati 781039, Assam, India
e-mail: sonali@iitg.ac.in

B. Venkatesh
e-mail: venkateshcjc@gmail.com

N. Bala Sai Krishna
e-mail: nadella.balasaikrishna@gmail.com

unlicensed users consist of multiple secondary base stations (sec-BSS) and secondary user terminals (sec-UTs). In the underlay mode of CRNs, prim-BS, prim-UT, sec-BS, and sec-UTs simultaneously use the spectrum by adjusting the power of sec-BS in a manner that interference to prim-UT is within a predefined limit [1]. The interference caused to the prim-UT depends on multiple factors, e.g., number of sec-BS, transmit power of each sec-BS, signal strength received at the prim-UT from the prim-BS, and channel conditions. Therefore, power allocation for each sec-BS is a challenging task keeping the interference of prim-UT below a threshold.

To allocate optimal power to sec-BSS there are centralized schemes, in which a central entity decides power for each sec-BS based on the collected channel condition between sec-UTs and prim-UT and transmitted power from prim-BS [2, 3]. The centralized schemes can provide an optimal solution, but it requires global information. Scalability is another issue with such schemes. As the number of sec-BS-src-UT pair grows, the data processing becomes complex and time-consuming. The CRNs are very dynamic in nature. Secondary users may come and go as per the user's needs. By the time a solution is obtained by the central entity, the network may be reconfigured. To address the scalability issues in underlay CRNs, distributed algorithms are found to be more useful. In this, the challenge is to find a near-optimal power allocation for sec-BSS without collecting channel and power statistics of other sec-UTs and prim-UTs.

In general, in a cognitive radio network, the surrounding radio elements keeps altering due to the mobility of users, interference issues, and the broadcast issues of the channels. Distributed approaches for power control are scalable with user growth [4], but the challenge is to obtain a near-optimal solution without having global information. In addition, frequent changes in CRN poses a further challenge of fast converging solution. Game theory is a composition of powerful mathematical models that investigate the strategic outcomes of multiple users. Many problems in wireless communications can be solved using various game theoretic approaches [5]. In game theory, The power allocation problem is modeled as a game where the sec-BSS are the players and every player would try to maximize its utility function with certain interference constraints from the prim-UTs [6, 7].

In this paper, we consider a scenario with numerous sec-BSS and different prim-UTs. Every sec-UT is served by only one sec-BS. The solution concept and problem formulation is based on a noncooperative coupled constraint game that tends to optimize the target variables (power) by checking for the existence of Lagrangian multipliers. A distributed algorithm has been proposed for the noncooperative power allocation scheme. Optimal power allocation for battery operated secondary users are important to improve their lifetime. In addition, in reality, batteries themselves are affected by multiple factors, including temperature, charging-discharging rates, and leakage. We consider a case of a realistic CRNs where the sec-UTs operate with an imperfect battery having some battery leakage. In such scenarios, the optimal transmit power must be chosen to prevent the battery from losing its energy such that network life improves. The proposed solution very well adapts to the dynamic nature of the CRN.

The paper organization is as follows. The network scenario considered in this work is discussed in Sect. 2. The distributed algorithm is discussed in Sect. 3. Numerical results the proposed algorithm are described in Sect. 4 and conclusions are listed in Sect. 5.

2 System Model

We consider a cognitive radio system comprising M prim-UTs and K sec-BSSs. Every sec-UT is served by a solitary sec-BS while one prim-BS serves many prim-UTs as the inclusion region of prim-BS is substantial and the inclusion zone of sec-BS is little. The main supposition made by us in regard to the conveyance of sec-BSSs is that every sec-UT is arranged near its master sec-BS and sec-BSSs don't cause any interference at nearby sec-UTs (Fig. 1).

Let's define some vectors and notations for formulating the problem.

- $pow = (p_1, p_2, p_3, \dots, p_k)$ where k is the number of sec-BSSs.
- $g = (g_1, g_2, g_3, \dots, g_k)$ where g_i is the channel gain between sec-BS i and served sec-UT i .
- $g_m = (g_m^1, g_m^2, g_m^3, \dots, g_m^k)$ where g_m^i is the channel gain between sec-BS i and prim-UT m with $m = 1, \dots, M$.
- $g_f = (g_1^f, g_2^f, g_3^f, \dots, g_k^f)$ where g_i^f is the channel gain between sec-BS i and sec-UT f where $i \neq f$.

As our framework is distributed, we don't have the channel gain between sec-BSSs and prim-UTs. This gain vector can be evaluated effectively by sending a pilot signal from prim-UT and detecting the sent pilot signal. To keep the QoS of prim-UTs high, the

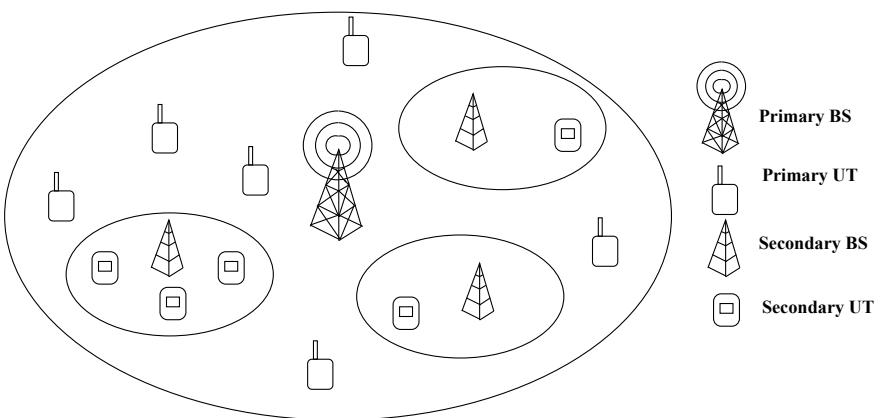


Fig. 1 Block diagram of the network scenario. The range of sec-BSSs is represented by the smaller circles and the range of the prim-BS is represented by the larger circle

most essential limitation is that the aggregate interference caused by sec-BSSs ought not surpass the threshold interference level.

Interference or the price paid by sec-BSSs at prim-UT m is

$$I_m = \text{pow}^T g_m \quad (1)$$

where $m = 1, 2, \dots, M$.

In an interference channel, the output signal vector \mathbf{y} is

$$\mathbf{y} = \mathbf{g}\mathbf{x} + \mathbf{w} \quad (2)$$

where \mathbf{x} is the input signal, \mathbf{g} is the channel gain matrix and \mathbf{w} is the noise vector. The signal to interference and noise ratio is defined as

$$\text{SINR} = \frac{P_k g(k, k)}{\sum_{i \neq k}^n P_i g(i, k) + \sigma^2} \quad (3)$$

where $g(i, k)$ is the channel gain from user i to sec-BS k . $\sum_{i \neq k}^n P_i g(i, k)$ is the interference caused by other sec-BSSs to the sec-BS k . Thus, in our scenario, SINR at a sec-UT f is

$$\text{SINR}_f = \frac{\text{pow}_f g_f}{\text{pow}_{-f}^T g_f + \sigma^2} \quad (4)$$

where pow_{-f}^T is the power vector assigned to sec-UTs excluding the f th user terminal. This is a general model where the g^f is not negligible. In our scenario, the g^f is negligible since the number of secondary networks we deal with is small, and these networks have a smaller coverage area. Thus, SINR reduces to

$$\text{SINR} = \frac{\text{pow}_f g_f}{\sigma^2}. \quad (5)$$

Now let's enforce the threshold interference constraint. Thus, for $m=1, 2, \dots, M$, the constraint is

$$\text{pow}^T g_m \leq I_T. \quad (6)$$

3 Problem Formulation and Solution Concept

To optimize secondary user power such that it does not violate the prim-UT's interference limits, the power allocation game must be formulated first.

$$\mathbf{Game} = \left\{ \mathcal{K}, \beta, \{u_f(p)\}_{f \in \mathcal{K}} \right\}$$

where the elements of the game are

1. Set of players: $\mathcal{K} = \{1, 2, \dots, K\}$.
2. The set containing the strategies: $\beta = \{p \mid p \in (0, p_{\max})\}$ and $p_T * g_m \leq I_T$ where $m = 1, 2, \dots, M$.
3. Set of utilities: The functions $u_f(p)$ where they are concave nondecreasing functions are defined as $\mathcal{U}_f(\gamma_f(p)) = \mathcal{U}_f\left(\frac{pow_f g_f}{\sigma^2 + pow_f^T g_f}\right)$.

3.1 Existence of Nash Equilibrium (NE)

In a strategic game, where the players choose deterministic strategies, the game $\{\mathcal{K}, \beta, \{u_f(p)\}_{f \in \mathcal{K}}\}$ has a NE, if for all $f \in \mathcal{K}$, the strategy set β for all players is non-empty compact subset of S (Euclidean space) and the continuity and concavity of the utility set is satisfied on β .

3.2 Uniqueness of Nash Equilibrium

The game will have a unique NE only in some special cases. When the payoff function of every player in the game satisfies strictly convex property and the feasible region should also satisfy the convex property, then there will be a unique NE.

3.3 Potential Game

A game $\{\mathcal{K}, \beta, \{u_f(p)\}_{f \in \mathcal{K}}\}$ to be a potential game, if there exists a function $F : \beta \rightarrow R$ which satisfies any one of the two following conditions

- (1) $F(x_j, x_{-j}) = u_j(x_j, x_{-j}) - u_j(x'_j, x_{-j})$ for any $j \in \mathcal{K}$, $x \in \beta$, and $x_j \in K_j$.
- (2) $\text{sgn}(F(x_j, x_{-j})) = \text{sgn}(u_j(x_j, x_{-j}) - u_j(x'_j, x_{-j}))$ for any $j \in \mathcal{K}$, $x \in \beta$, and $x_j \in K_j$, where $\text{sgn}(\cdot)$ is the signum function.

If such a function exists it is called a potential function. If the potential function F satisfies the first or second condition then the game can be categorized as an exact potential game or an ordinary potential game, respectively. These two conditions imply that each individual player's advantage is subject to the next part player's enthusiasm for the gathering. If all players in a potential game take optimal choices or better strategies one after another in a series then it will end up with NE (which maximizes the value of the potential function F) infinite steps.

3.4 Normalized Nash Equilibrium

Consider a game $\{\mathcal{K}, \beta, \{u_f(p)\}_{f \in \mathcal{K}}\}$, where the elements of the game are

1. Set of players: $\mathcal{K} = \{1, 2, 3, 4, \dots, K\}$.
2. The set containing the strategies: $\beta = \{p \mid p \in (0, p_{\max})\}$ and $p^T * g_m \leq I_T$ for $m = 1, 2, \dots, M$.
3. Set of utilities: The functions $u_f(p)$, where they are concave nondecreasing functions, are defined as follows.

We need to choose to transmit power of each of the K sec-BSs in a game G so that it will improve the utilization or overall payoff of the sec-BSs, i.e., we need to apply a good power allocation policy for efficient utilization. If we consider the power allocation vector pow^* as NE then it should satisfy the following conditions

For every $k \in \mathcal{K}$ and pow_k such that

$$(pow_1^*, \dots, pow_{k-1}^*, pow_k^*, pow_{k+1}^*, \dots, pow_K^*) \in \beta$$

$$\mathcal{U}_f\left(\frac{pow_f^* g_f}{pow_{-f}^{*T} g_f + \sigma^2}\right) \geq \mathcal{U}_f\left(\frac{pow_f g_f}{pow_{-f}^{*T} g_f + \sigma^2}\right)$$

By observing that each $\mathcal{U}_f\left(\frac{pow_f^* g_f}{pow_{-f}^{*T} g_f + \sigma^2}\right)$ is continuous in R_+^K and the strategy set β is closed, convex, and bounded, we can use sufficient Karush-Kuhn-Tucker (KKT) conditions.

If pow^* is a NE in β then there exist K vectors $\lambda^k = (\lambda_1^k, \lambda_2^k, \dots, \lambda_M^k)$ with $\lambda^k \geq 0$ such that pow^* satisfies the following equation for $k = 1, 2, \dots, K$ and $m = 1, 2, \dots, M$

$$\lambda_m^k (pow^T g_m - I_T) = 0.$$

We now formulate a distributed algorithm based on NNE for the power allocation scheme [8]. The cost of interference due to player k at prim-UT m is analogous to the Lagrangian multiplier [9] λ_m^k . Hence, in an NNE the primary users need not select different costs for the players. As an initial step, every λ_m^k is chosen randomly. As the iterations proceed, we update power and lambda.

Algorithm 1 Distributed Power Allocation

Choose initial cost λ_m^k randomly, due to player k at prim-UT m .

Update:

1. Power set for every sec-BS k is given by
 $pow_k^i = \text{argmax}(\text{potential function}) - pow_k * \lambda^m T * g^k$.
 2. The cost set by prim-UT m is $\lambda_m^{i+1} = (\lambda_m^i + \partial(h_k * pow_k^i - I_T))$.
-

4 Numerical Results

In this section we obtain the simulation results with the following parameters. Initially, power values for each sec-BS will be allocated between 0 and 1 mW. The maximum transmission power for each sec-BS is 1 mW and the maximum interference level for each prim-UT is 1 mW. The minimum SNR for each sec-UT is 1 dB. We have allocated channel gain between sec-BS and prim-UT between 0 and 1. We have simulated for K sec-BSSs and M prim-UTs by first considering without and then with battery leakage.

4.1 Without Battery Leakage

Figure 2 shows the transmission power values of secondary users vs iteration number for $K = 5$ and $M = 3$. Initially, sec-BSSs got random power values for transmission between 0 and 1 mW and after four iterations there is no change in power values of each secondary user. This implies we have attained an equilibrium state where the overall utility of secondary users is maximized and hence power values are converging. Power values have converged after four iterations.

In Fig. 3, for $K = 10$ and $M = 5$, power values have converged after three iterations. This shows that our algorithm is fast converging and suitable for dynamic CRNs.

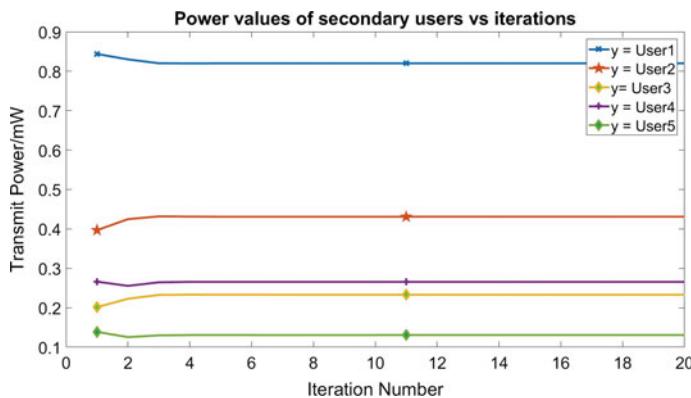


Fig. 2 Transmission power values for five sec-BSSs and three prim-UTs

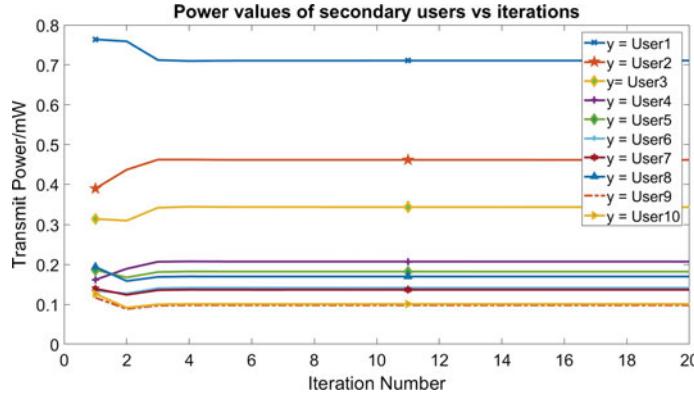


Fig. 3 Transmission power values for 10 sec-BSSs and five prim-UTs

4.2 With Battery Leakage

Providing optimal power management is a challenge, particularly, for secondary users equipped with a battery that provides limited energy. Also, we consider the battery leakage while allocating the power to sec-BS.

Initial energy for sec-BSSs are given as 1000 mJ. The battery leakage value for sec-BS $k(k = 1, 2, \dots, K)$ have been allocated k mJ/s. Each sec-BS has been allocated transmit power between 0 and 1 mW initially, in a manner that sec-BS having higher battery leakage gets allocated lesser power and vice versa, in order to increase the life of the network.

Figure 4 shows the transmission power values of SUs vs iterations for $K = 10$ and $M = 5$. The power values have converged after three iterations and later User10 became inactive and again power values have been allocated accordingly. In the

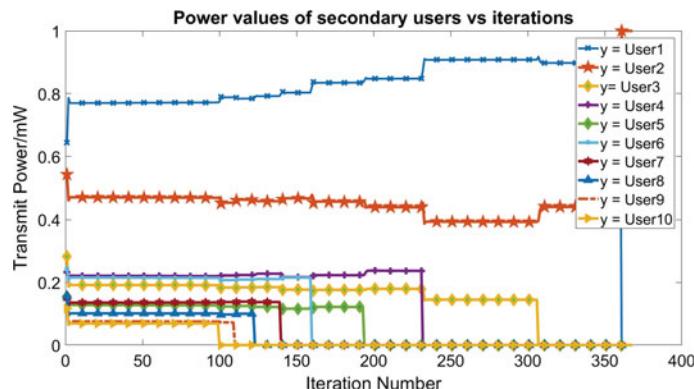


Fig. 4 Transmission power values for 10 sec-BSSs and five prim-UTs with battery leakage

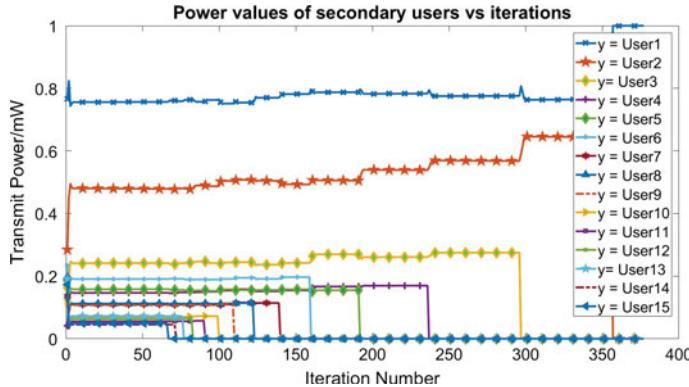


Fig. 5 Transmission power values for 15 sec-BSSs and eight prim-UTs with battery leakage

figure, User10 has become inactive after 100 iterations because User10 is having battery leakage 10 mJ/s and later User nine becomes inactive and so on. Average lifetime has been increased. When a less number of users are remaining, the algorithm adjusts the power and more power is given to the remaining users. This shows the utility of the proposed solution in a dynamic CRN.

Figure 5 shows the transmission power values of sec-BS versus iterations for $K = 15$ and $M = 8$, with battery leakage. User15 has become inactive after 60 iterations because User15 is having the highest battery leakage of 15 mJ/s and later User14 becomes inactive and so on. Our algorithm converges fast in this case as well and allocates power in a manner that average network lifetime increases.

5 Conclusions

We have presented a noncooperative game theoretic solution to the power allocation problem in the CRN. It is found that the power allocation satisfies Normalized Nash equilibrium and also fast converging solution. While allocating the power, sec-UT's interference limit was taken into account. The method was applied to battery limited CRNs while taking battery leakage into account. This is a near practical scenario. The proposed algorithm works well and provides power allocation so that average network life is increased. For CRN with changing number of sec-BSSs, algorithm dynamically adjusts the power allocation while maintaining the interference to the prim-UT within a prescribed limit. This maximizes the secondary user's throughput by allocating the maximum possible power.

References

1. A. Goldsmith, S.A. Jafar, I. Maric, S. Srinivasa, Breaking spectrum gridlock with cognitive radios: an information theoretic perspective. Proc. IEEE **97**(5), 894–914 (2009)
2. A. Tsakmalis, S. Chatzinotas, B. Ottersten, Centralized power control in cognitive radio networks using modulation and coding classification feedback. IEEE Trans. Cognit. Commun. Netw. **2**(3), 223–237 (2016)
3. Z. Chen, F. Gao, Cooperative-generalized-sensing-based spectrum sharing approach for centralized cognitive radio networks. IEEE Trans. Veh. Technol. **65**(5), 3760–3764 (2016)
4. S. Parsaeefard, A.R. Sharafat, Robust distributed power control in cognitive radio networks. IEEE Trans. Mob. Comput. **12**(4), 609–620 (2013)
5. Z. Han, D. Niyato, W. Saad, T. Basar, A. Hjorungnes, *Game Theory in Wireless and Communication Networks: Theory, Models and Applications* (Cambridge University Press, 2011)
6. P. Zhou, Y. Chang, J.A. Copeland, Reinforcement learning for repeated power control game in cognitive radio networks. IEEE J. Select. Areas Commun. **30**(1), 54–69 (2012)
7. G. Yang, B. Li, X. Tan, X. Wang, Adaptive power control algorithm in cognitive radio based on game theory. IET Commun. **9**, 1807–1811 (2015)
8. A. Ghosh, L. Cottatellucci, E. Altman, Normalized nash equilibrium for power allocation in cognitive radio networks. IEEE Trans. Cognit. Commun. Netw. **1**(1), 86–99 (2015)
9. Q. Zhu, A Lagrangian approach to constrained potential games: theory and examples, in *IEEE Conference on Decision and Control* (2008), pp. 2420–2425

Relay Selection-Based Physical-Layer Security Enhancement in Cooperative Wireless Network



Shamganth Kumarapandian and Martin James Sibley

Abstract Broadcast nature during the data propagation and wireless transmission from the source to destination node can be easily overheard by the unauthorised users due to security issues. It cause interception and is highly vulnerable to eavesdropping effect. In this paper, a hybrid algorithm is proposed to overcome the limitations in physical-layer security and achieve optimal local solution. Cooperative-based relay selection approach is proposed to enhance the network range and durability in wireless communication and double threshold-based relay selection scheme to improve the spectral efficiency and overall quality of the communication system. Furthermore, the hybrid evaluation algorithm enhances the performance parameters such as signal strength and channel capacity; also it minimises the number of nodes. The proposed relay selection scheme is compared with direct transmission, P-AFbORS, P-DFbORS schemes and it is observed that better results are achieved from the proposed multi-relay selection scheme as compared to other the existing relay selection schemes.

Keywords Cooperative wireless communications · Threshold-based relay selection · Hybrid evaluation algorithm · Physical-layer security

1 Introduction

Physical-layer security has proven to be an effective alternative for secure data transmission, and it can be achieved by modifying the capacity of the main link channel, i.e. from source to destination higher than wiretap link, i.e. from source to the eavesdropper [1]. During this process, there is a chance of security issues, if the rate of secrecy capacity falls to zero. This is generally observed due to fading effect and several works have been reported in the literature to enhance the rate of secrecy capacity such as the multiple numbers of antennas and cooperative relays [2, 3]. The physical-layer security techniques deployed with various diversities are as follows.

S. Kumarapandian (✉) · M. J. Sibley

School of Computing and Engineering, University of Huddersfield, Huddersfield, UK
e-mail: shamganth@gmail.com

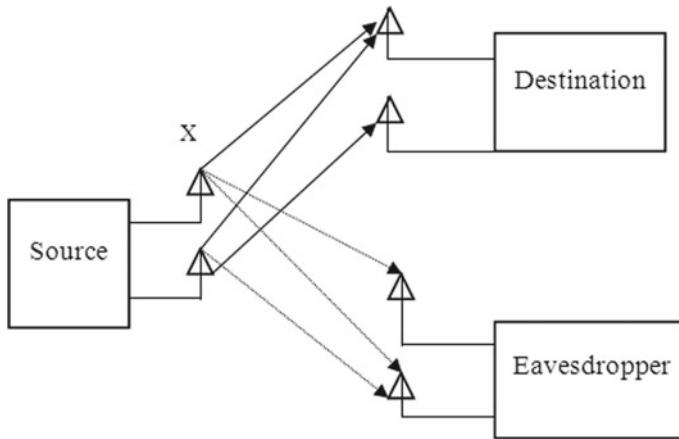


Fig. 1 MIMO system with a single source node and destination [4]

1.1 *MIMO Diversity*

Figure 1 shows the Multiple-Input Multiple-Output communication (MIMO) system along with the eavesdropper effect. The notation X , Y_d and Y_e in Fig. 1 represents the antennas at the source node, destination node and the eavesdropper. It is observed that MIMO structure can be exploited through eavesdropper effect to enhance the wiretap channel capacity amongst source to destination. Thus, improper design can increase the rate of fraud in wireless transmission. Initially, the destination node accesses the data from the main channel matrix H_x followed by decoding process through space-time with the estimation of \hat{H}_x which leads to capture diversity gain from the main channel. Further, the eavesdropper node can also calculate the wiretap channel matrix H_{wc} and diversity through corresponding space-time decoding algorithm. Hence, the traditional MIMO process is not found to be effective against the eavesdropping effect.

1.2 *Diversity Through Multiuser*

Figure 2 shows the system model of multiuser diversity communication system for enhancing the physical-layer security. The base station serves multiple mobile users. The communication amongst the receiver node and the transmitter node is achieved through Time Division Multiple Access (TDMA) technique and Orthogonal Frequency Division Multiple Access (OFDMA) technique. In traditional systems, the user with high throughput is considered to access the OFDM subcarrier to achieve maximum transmission capacity. This process relies on the knowledge of the key

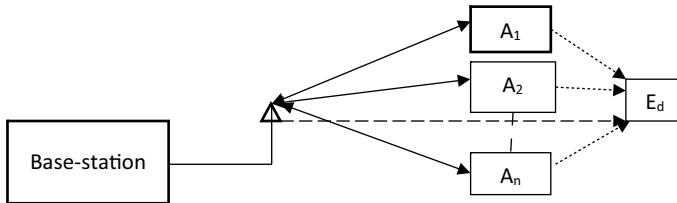


Fig. 2 Diversity in multiuser wireless networks [5]

channel information. But there exists a limitation of deep fading and propagation loss, if the user is farthest from the base station.

1.3 Cooperative Diversity System

Figure 3 shows the system model of the single source–destination cooperative diversity system along with the eavesdropper effect. The relays A_1, A_2, \dots, A_n are deployed between source and destination to assist and enhance the rate of signal transmission. The signal transmitted from the source node is forwarded by N -relays to the destination node. In general, relaying protocols used at the relays are Amplify-and-Forward (AF) and Decode-and-Forward (DF). The signal is broadcasted by the source node to the relay node and the retransmission of signal from relay node to the destination. Individual transmission is vulnerable to the eavesdropping attack and additional care should be taken during the design of transmission. Cooperative beamforming is used to increase the number of relays and enhance channel capacity. From the above-mentioned study on cooperative diversity system, it is observed that the eavesdropper effect can be reduced by increasing the strength of the received signal at the destination node and by choosing best relay selection algorithms.

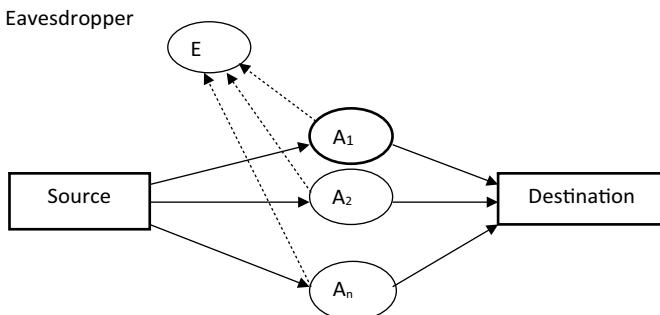


Fig. 3 Single source and destination cooperative diversity system [6]

2 Prior Related Research

There is a lot of research interests shown in relay selection techniques along with physical-layer security are ascertained to enhance the performance of the system. The existing research is as follows.

Physical-Layer Security (PHY) plays an active role in providing security and its application on cooperative relaying is used to enhance the network range and durability. The relay nodes exploit the properties of the physical layer in wireless channels and provide assistance to secure transmission from the source node to the destination. Further, optimal sequential deployment is analysed, and a new measurement-based optimal technique for the dual-hop wireless relay network is developed by Ghosh et al. [7]. The author has induced multi-connectivity approach in the individual system by providing effective communication amongst individual node and the neighbouring node. Markov decision process is considered to evaluate the problem related to cost objective, routing and placement, and numerical exploratory approach is used to evaluate the developed design. It is ascertained from the study that average outage per relay for the value $c = 1$ and $c = 0.1$, gives a high end-to-end outage and for $c = 10$, provides less and practical end-to-end outage. The experimental analysis also shows approximately equal values to the ratio of per setup cost to the simulated per setup cost. Furthermore, a shadow fading model on the basis of the measurement technique for the application of vehicle-to-vehicle simulation network is proposed by Abbas et al. [8]. The propagation channel for the application of vehicle-to-vehicle network has key importance during designing and has a direct impact on the performance of Vehicular Ad Hoc Networks (VANETs). The author has developed the model on the basis of real-time scenario in the urban areas and the highways. Measurement data considered for the research is divided into three categories namely Obstructed Line-Of-Sight (OLOS), Line-Of-Sight (LOS) and the online sight arise due to buildings, and an approach has been developed in incorporating the LOS/OLOS model into the VANET-based simulators. From the study, it is observed that if the distance amongst the two nodes is 300–400 m, there is a 60% no packet loss amongst successful nodes and for the distance range of 400 m under LOS/OLOS model, better packet reception ratio along with high interference is achieved.

Likewise, relay selection technique for cooperative relays on the basis of performance analysis for finite energy storage wireless transmission is developed by Liu et al. [9]. The interruption of data arises due to the relay selection and an effective timing structure to enable the relay selection along with the channel state information, and the power status of the relay is considered for the analysis. The author has developed the finite-state Markov model to analyse the evolution of power consumption in batteries and derived the resultant probability. Simulation analysis is considered for evaluation and results show that bigger size battery with high capacity is found to be a beneficial infinite energy storage of EH relays. Furthermore, scrutiny shows that two factors are responsible for variable performance and they are a set of potential relays and selection of overhead for RS information collection. The solution can be

achieved through threshold-based relay selection approach by minimising the number of competing relays and channel estimation overhead. Swain [10] has studied the effects of IEEE 802.16j MMR WiMAX network architecture along with its computational complexity and signal processing operation. Threshold-based max–min relay selection technique along with harmonics parameter is considered to improve the performance of a conventional amplify–forward process and the decode–forward process for the assistance of multi-relay selection in IEEE 802.16j MMR WiMAX network. Further, several standard diversity combinational techniques such as selection combination and maximal ratio combination are considered for signal transmission and reception. The results obtained from the study show that threshold-based harmonic mean SNR relay selection technique provides better performance results in terms of SER compared to other existing relay selection techniques. Furthermore, improved channel capacity is achieved compared to the individual harmonic mean algorithm. In our earlier work, a novel selection scheme comprising a twofold threshold technique in the cooperative network is proposed by S.Kumarapandian [11]. The combinational approach of Dual-Threshold MRC (DT-MRC) along with differential AF modulation is developed, and the analysis is done in terms of existing path estimators and active branches for a reduction in power consumption. Statistical characterization is considered to evaluate the proposed relay selection scheme. Initial functional parameters such as cumulative distribution function, probability distribution and moment-generating function are derived for the combinational SNR of developed relay selection scheme. From the study, it is observed that there is a reduction in the average channel estimation with an increase in the normalised SNR value.

A novel adaptive relay selection is developed to analyse the variations in the conditions of the channel and to enhance the PER value and the relaying probability rate [12]. The author has done several studies on the impact of node density, size of the data packet, coherence time, path distance amongst source and destination and spatial reusability. The developed protocol is analysed using cooperative relay communication in the wireless medium. The experimental analysis is done in terms of several parameters such as channel coherence time, a density of the node and size of the data packet. The results obtained from the study provides improved performance in terms of throughput and reliability. By considering security aspects during the transmission of data, a novel cooperative beam-forming technique comprising physical-layer security and partial relay selection is developed by Qian et al. [13]. The author has considered source–destination pair and number of decode–forward relays along with the effect of the eavesdropper and two relay strategies namely linear-complexity relay ordering and exponential complexity exhaustive search for the analysis. The study on the performance analysis shows that the developed scheme provides high secrecy capacity compared to existing search schemes. An enhanced relay selection scheme comprising Proposed Source–Relay Selection (PSRS) along with physical-layer security is implemented by Shim et al. [14]. The author has investigated the privacy performance of the opportunistic scheduling in multi-relay cooperative modelling which ascertains the improvement in physical-layer security. The combined approach is comprising Maximal Ratio Combing (MRC) along with

Selection Combining (SC) to analyse the effect of the eavesdropper and the results obtained are calculated in terms of Secrecy Outage Probability (SOP). From the study, it is observed that the developed technique achieves higher SOP rate compared to PSRS.

From the aforementioned study, it is observed that still there exist limitations in terms of security and optimal solution. The high secrecy can be achieved with a minimal number of relays and efficient relay selection techniques. In this research, an optimal dual-threshold-based relay selection technique is considered along with optimal relay assignment to evaluate eavesdropper effect.

3 Proposed Methodology

3.1 Double Threshold-Based Relay Selection Scheme

The main idea of the proposed relay selection scheme is to minimize the channel estimation and reduce power consumption by selecting an optimal node. The proposed scheme is designed in such a way to enhance the lifetime of the relay node. The limitation arises due to the single threshold relay selection scheme; [15] is considered during the analysis and it is minimised by proposing a double threshold relay selection scheme. Also, combining the threshold-based relay selection with the optimal relay selection scheme enhances the physical-layer security. The mode of operation of the proposed scheme is shown below:

Step 1: Initially, the value of ‘n’ which defines node number is set to zero and the threshold SNR value Γ_c is set to zero.

Step 2: The direct transmission of the source to destination node takes place in this step. The threshold SNR at the destination is set to Γ_{sd} .

Step 3: Initially, the relay node is selected. The data from the source node is received by the relay node, and later the data is forwarded by the relay to the destination. The threshold SNR value is set for the relay and the same is considered for transmission which is given by $\Gamma_{s,m}$.

If the number of relay is equal to the last relay node, the process is terminated, and the last relay node is selected as an optimal node.

Step 4: The minimum predefined threshold SNR (Γ_{IT}) is fixed, and the same is tested with the SNR at the relay value. If the SNR at the relay exceeds the predefined threshold value, that relay node is selected as an optimal relay. If it fails, then the next relay is selected, and the process continues until it reaches the final relay.

Step 5: In the next stage, the maximum value of threshold SNR with respect to relay node (α_1) and the maximum of predefined SNR with respect to relay node and destination node, and the source node to destination node (α_2) is calculated.

Step 6: If the selected relay exceeds the output threshold, that relay is selected as an optimal relay. If it does not exceed, then the process continues.

In this model, two threshold values namely input threshold and output threshold are considered for analysis. The input threshold is considered to evaluate the quality of the selected relay and the output threshold for evaluating the overall quality of the communication system. The spectral efficiency is enhanced since the relays above the input threshold is in the listening mode and the remaining relays will be in silent mode.

3.2 Optimal Relay Selection Schemes

The optimal relay selection scheme is considered to enhance the physical-layer security of the communication system through AF and DF relaying protocols. Further optimal relay selection schemes such as P-AFbORS and P-DFbORS [16] are defined for the evaluation purpose. The detailed description of the proposed schemes are as follows.

3.2.1 Amplify-and-Forward Scheme (AF)

In this study, the AF relaying protocol is used. The relay node retransmits an amplified version of the data to the destination [12]. Through this process, the optimal relay node is selected for effective transmission. The total amount of transmission power amongst the source and the relay is limited to value ‘P’ to provide fair comparison with direct data transmission. Furthermore, the equal power allocation is used to calculate the transmitted power amongst source and the relay which is given by $p/2$. The received signal at the relay q_i is calculated on the basis of assumption that the source node transfer signal S with the power $p/2$ is given by

$$q_i = \sqrt{\frac{p}{2}} h_{si} S + t_i \quad (1)$$

where

h_{si} Fading coefficient observed from source to q_i .

t_i AWGN at q_i .

3.2.2 P-AFbORS

The P-AFbORS protocol is used in this research to select the optimal relay and to enhance the capacity of AF relaying transmission [16]. Optimal relay selection criteria on the basis of AF relaying

$$\text{OR} = \arg \max_{i \in S} C_i^{AF} \quad (2)$$

where S denotes the multiple relay set.

The above equation comprises of both wiretap links and the main links. Furthermore, distributive or centralised relay selection scheme is implemented from the proposed criteria for optimal relay selection. P-AFbORS also considers main links and the wiretap links and the equation is given by [16]

$$\text{OR} = \arg \max_{i \in S} \frac{1 + \frac{|H_{SI}|^2 |H_{ID}|^2 P}{2(|H_{SI}|^2 + |H_{ID}|^2) \sigma_N^2}}{1 + \frac{|H_{SI}|^2 |H_{IE}|^2 P}{2(|H_{SI}|^2 + |H_{IE}|^2) \sigma_N^2}} \quad (3)$$

where S defines the set of 'M' relays. From Eq. 3, it is observed that main links CSI $|H_{SI}|^2$ and $|H_{ID}|^2$ and wiretap link CSI $|H_{IE}|^2$ are considered for the evaluation. Transmit power 'P' is initially defined as the known parameter. The noise variance σ_N^2 is given by

$$\sigma_N^2 = kTB \quad (4)$$

where

- T - Absolute room temperature (290 K)
- k - Boltzmann constant, i.e. $1.38 * 10^{-23}$ J/s
- B - System bandwidth.

3.2.3 Decode-and-Forward Scheme (DF)

In decode-and-forward relaying scheme, the received data obtained from the source node is decoded at the initial stage followed by re-encoding and transmission of the signal to the destination. Subsequently, the multiple number of relays accepts the transmitted signal obtained from the source node and an attempt is made on decoding at the same. The optimal relay node is considered for re-encoding and transmission of decoded signal. The two-hop DF transmission provides failure results, if there is an existence of failure amongst any one of the source-relay or the relay-destination path. Let R_i be the optimal relay considered for calculating the DF transmission capacity amongst the source to destination and it is given by [16],

$$C_{sid}^{DF} = \min(C_{si}, C_{id}) \quad (5)$$

where

- C_{si} Channel capacity from source to R_i
- C_{id} Channel capacity from R_i to destination.

The relay which enhances the DF relaying capacity is considered as an optimal relay and equation for calculating the criteria for optimal relay is given by

$$\text{Optimal Relay} = \arg \max_{i \subseteq R} C_i^{DF} \quad (6)$$

From the above equation, it is observed that both wiretap and main link parameters are considered to calculate the optimal relay. The probability of interception of P-DFbORS scheme is calculated by using the concept of intercept event given by

$$P_{\text{intercept}}^{P_{\text{DFbORS}}} = P(\max_{i \in R} C_i^{DF})_0 \quad (7)$$

3.2.4 VMWMC Formulation

In this study, weighted bipartite graph for the cooperative wireless network is designed to analyse and formulate the problem of joint optimization and to clearly describe the correlation amongst the pair of transmission and relay nodes. A virtual relay node has been developed for representing the direct and cooperative transmission in a uniform conceptual way. Further, the problem has been analysed regarding variations in the maximum weight matching where the weight parameters are defined regarding power value function to meet the power constraints (VMWMC). The VMWMC is observed as a non-convex problem in which the complexity increases with an increase in the number of relays [17]. In this paper, an optimal solution for the VMWMC problem was calculated through an exhaustive search of exponential complexity $O(a^m)$ where 'a' defines the transmission pairs and 'm' represents a number of relay nodes.

3.2.5 Hybrid Evaluation Algorithm (HEA)

In this work, a hybrid evaluation algorithm is developed for optimal relay selection. The proposed algorithm is effective in solving the problems related to the condition namely 'single problem has multiple solutions'. This algorithm comprises the number of steps namely mutation and crossover to address the complexity issues. Crossover selection is defined as a technique of combining the genetic information of two individuals so that the coding can be appropriately selected. Several optimal cooperative communication parameters comprising signal strength, channel capacity and the number of nodes can be minimised through the proposed algorithm to reduce the cost amongst the nodes. The crossover method is used for optimal path selection and mutation operator is used to calculate the capacity amongst the path of the source to the relay node and to the destination node. During the data transmission, when the receiver accepts the packet of data, the request is forwarded to evaluation algorithm to compute the number of paths amongst source to destination and the selection of optimal data transfer path.

Furthermore, a probabilistic adaptive-based crossover and mutation is considered to overcome the limitation of getting the optimal local solution. The fitness evaluation is calculated in terms of time efficiency and accuracy, and elite solutions are computed through exact fitness.

4 Performance Evaluation and Simulation Results

The experimental analysis is conducted by initially considering two to eight relays. The iteration number is directly proportional to the number of relays, and it increases with increase in the relay count. The cooperative wireless network is designed with the number of nodes distributed in a sectional area of $500\text{ m} \times 300\text{ m}$. The tolerance level is set to 0.00000001 for obtaining precise simulation results. Total of 100 relays with three grids and a receiver is considered. Two test cases are discussed in which one consists of a random topology of the source node, destination and relay node deployed randomly in an area defined by the ad hoc network. The second comprises tree topology in which entire source node is communicated with the single destination or sink node. It is described as a relay cellular network. The simulation parameters of the proposed technique are as follows (Table 1).

Figure 4 provides the details regarding the techniques considered in the proposed relay selection scheme. Multiple numbers of nodes are found for the selection of optimal relay in multi-hop relay selection. The black points in the figure represent the relay node except the source node (node 1) and the destination node (node number 100) and the points with red colour represents optimal relays path from source to destination node. Amplify-and-forward technique is considered for the above analysis. Furthermore, the asymptotic intercept probability analysis is considered to enhance physical-layer security in multiuser cooperative networks. Asymptotic analysis is derived to improve the secrecy performance and to calculate the secrecy performance in terms of high Main-to-Eavesdropper Ratio (MER).

Figure 5 represents the rate of probability of interest over Main-to-Eavesdropper Ratio (MER). From the graph, it is observed that techniques such as P-AFbORS and P-DFbORS are better compared to direct transmission in terms of intercept

Table 1 Simulation parameters

S. no	Parameters	Values
1	Area	$500\text{ m} \times 300\text{ m}$
2	Number of nodes	100
3	Grid	3
4	Receiver	1
5	Tolerance	0.00000001

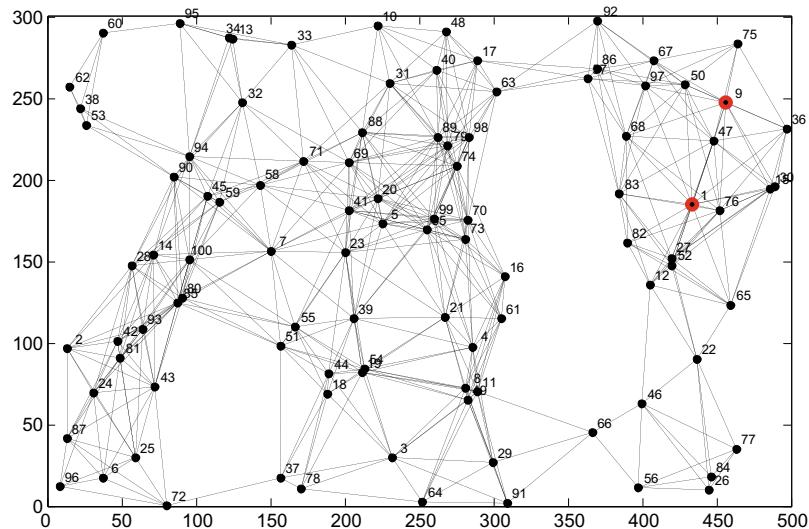


Fig. 4 Proposed relay selection techniques

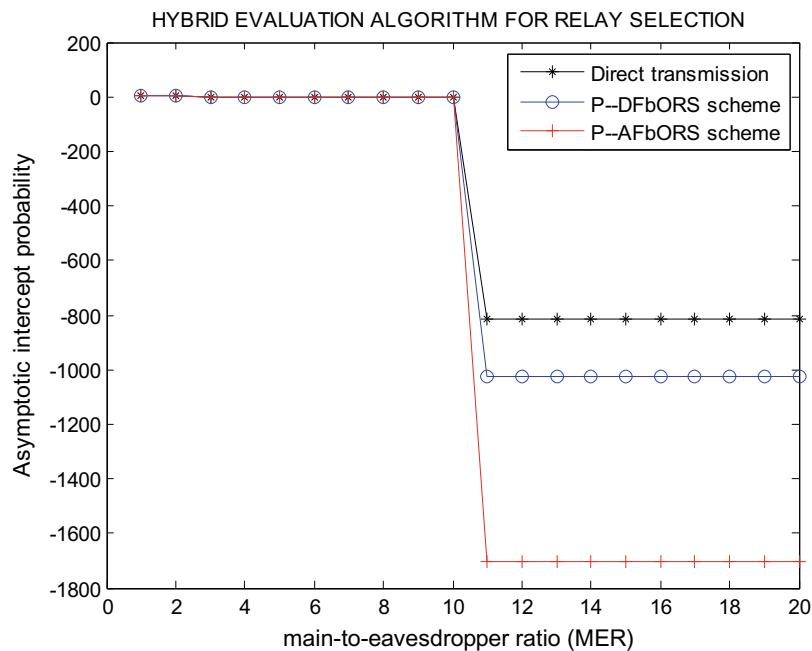


Fig. 5 Asymptotic intercept probability analysis

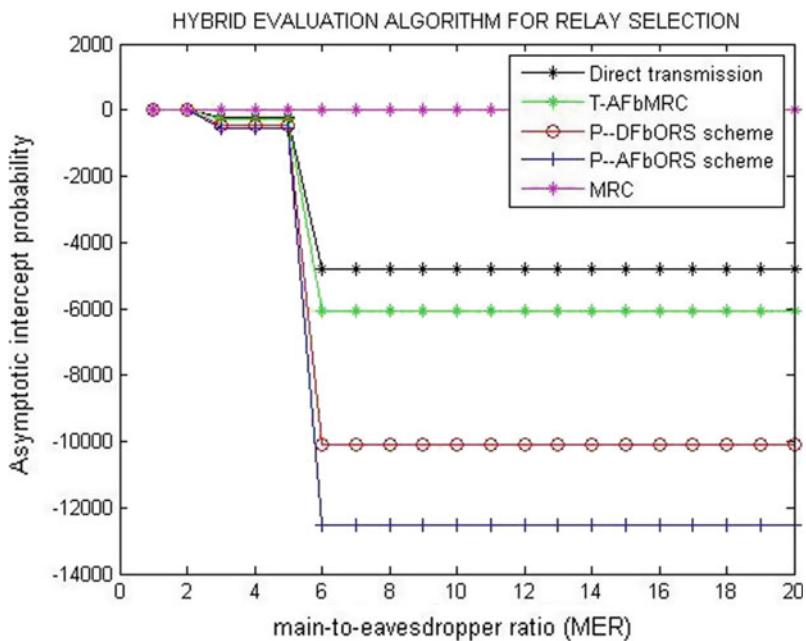


Fig. 6 Proposed performance analysis of Hybrid evaluation algorithm

probability and eavesdropping attack. Figure 6 compares the proposed scheme with the traditional Amplify and Forward based MRC scheme

Figure 7 provides the details regarding the number of messages versus node ID in terms of bar graph.

Figure 8 provides the details regarding the network capacity obtained from the hybrid evaluation algorithm. The network capacity defines the capability of the network link to transfer data from one node to another in a network. It is observed that through the proposed hybrid model, the network link capacity is increasing gradually with increase in the number of iterations.

Figure 9 provides the throughput graph obtained from the proposed evaluation algorithm. It is defined as the ratio of time taken by the receiver to accept the total amount of data from the source to the last data packet reaching the destination. The equation for the same is given by

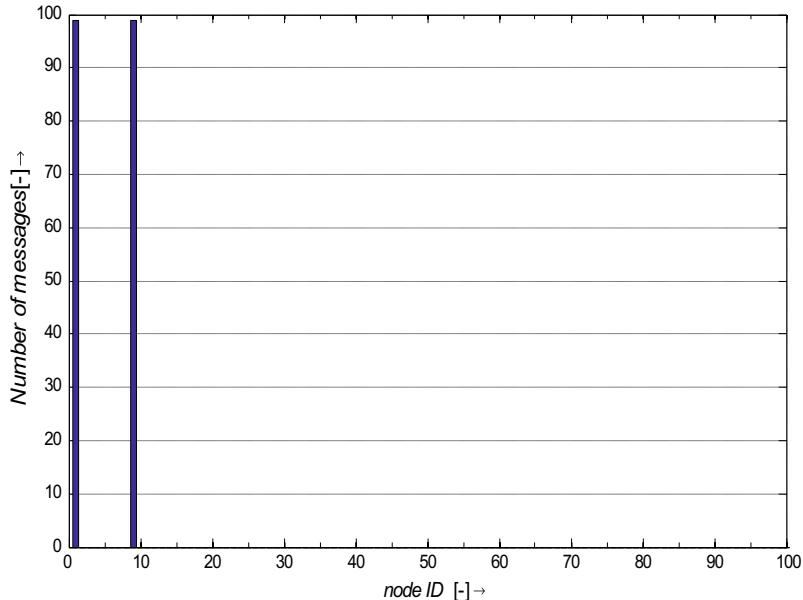


Fig. 7 Messages versus node ID

$$X = \frac{C}{T} \quad (8)$$

where

X Throughput

C Total number of request realised by the communication system

T Total time for examination of the communication system.

The accuracy is calculated in terms of network throughput and it is observed that the proposed algorithm has strong variation in the throughput which showcases loss of packet or packet drop. Further, it is shown that the throughput gradually increases with increase in number of iterations.

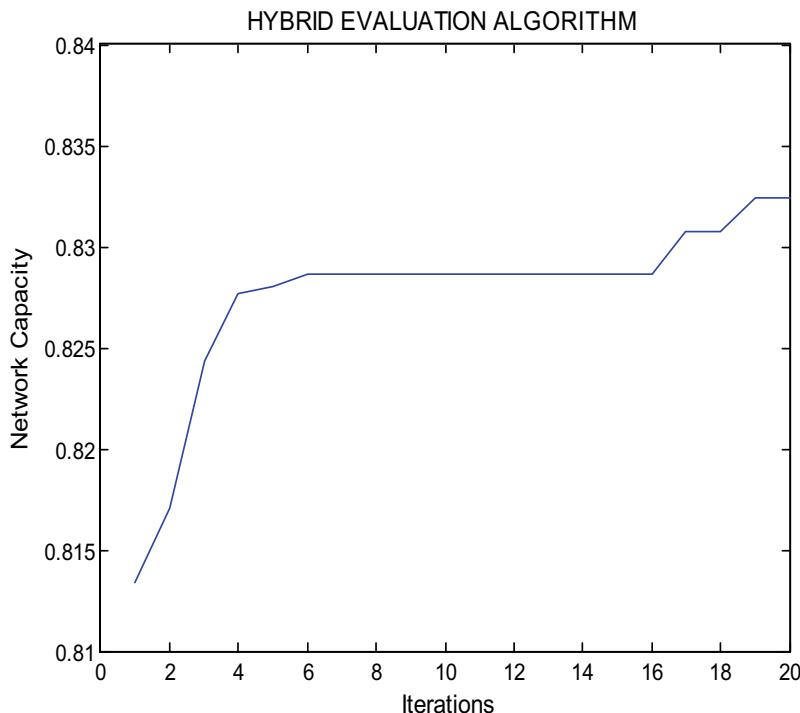


Fig. 8 Impact of the proposed scheme on the network capacity

5 Conclusion

In this research, a cooperative-based relay selection technique is proposed to enhance the rate of data transfer in a network, durability and physical-layer security in a wireless communication system. A novel double threshold-based relay selection technique is developed for the selection of optimal relay and to amplify and forward the data from the source node to the destination. The issues such as joint optimization constraint, power allocation and relay assignment are designed using VMWMC, and it is analysed in terms of asymptotic intercept probability analysis, network throughput and network capacity improvements through proposed novel HEA. Optimal relaying schemes namely P-AFbORS and P-DFbORS are applied to increase the physical-layer security against eavesdropping attack and it is inferred that the proposed multi-relay selection technique provides better results and throughput and it is directly proportional to the increase in a number of relays. This research is limited to one-way Cooperative wireless networks, and in future, the same can be implemented for full-duplex cooperative wireless networks.

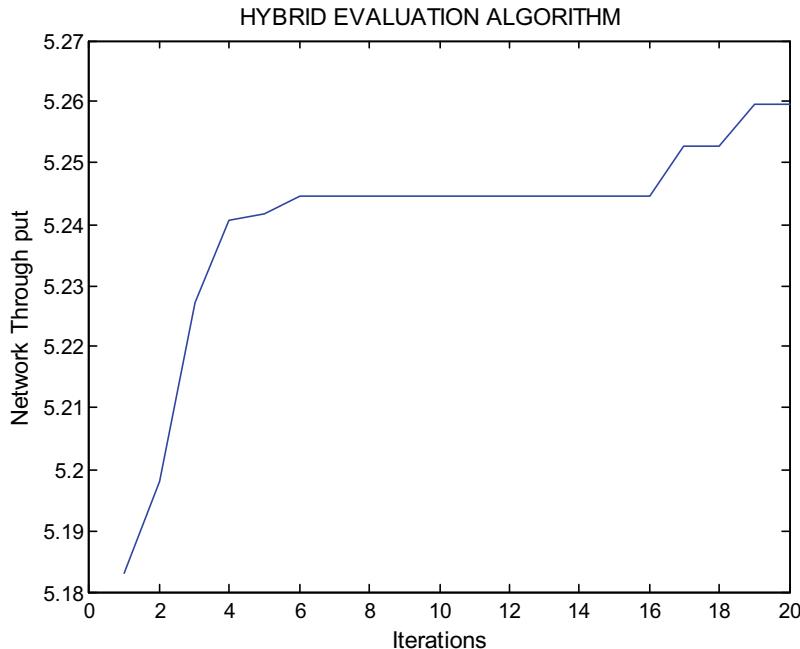


Fig. 9 Network throughput of hybrid evaluation algorithm

References

- Y. Zou, J. Zhu, X. Wang, L. Hanzo, A survey on wireless security: technical challenges, recent advances, and future trends. *Proc. IEEE* 1–39 (2016)
- X. Chen, C. Zhong, C. Yuen, H.H. Chen, Multi-antenna relay aided wireless physical layer security. *IEEE Commun. Mag.* **53**(12), 40–46 (2015)
- Q. Li, Y. Yang, W.K. Ma, M. Lin, J. Ge, J. Lin, Robust cooperative beamforming and artificial noise design for physical-layer secrecy in AF multi-antenna multi-relay networks. *IEEE Trans. Signal Process.* **63**(1), 206–220 (2015)
- S. Zhang, Z. Ying, J. Xiong, S. He, Ultrawideband MIMO/diversity antennas with a tree-like structure to enhance wideband isolation. *IEEE Antennas Wirel. Propag. Lett.* **8**, 1279–1282 (2009)
- J.L. Rebelatto, B.F. Uchôa-Filho, Y. Li, B. Vucetic, Multiuser cooperative diversity through network coding based on classical coding theory. *IEEE Trans. Signal Process.* **60**(2), 916–926 (2012)
- J.N. Laneman, D.N. Tse, G.W. Wornell, Cooperative diversity in wireless networks: efficient protocols and outage behavior. *IEEE Trans. Inf. Theory* **50**(12), 3062–3080 (2004)
- A. Ghosh, A. Chattopadhyay, A. Arora, A. Kumar, Measurement based as-you-go deployment of two-connected wireless relay networks. *ACM Trans. Sens. Netw. (TOSN)* **13**(3), 23 (2017)
- T. Abbas, K. Sjöberg, J. Karedal, F. Tufvesson, A measurement based shadow fading model for vehicle-to-vehicle network simulations. *Int. J. Antennas Propag.* (2015)
- K.H. Liu, Performance analysis of relay selection for cooperative relays based on wireless power transfer with finite energy storage. *IEEE Trans. Veh. Technol.* **65**(7), 5110–5121 (2016)

10. C.M.K. Swain, S. Das, Effects of threshold based relay selection algorithms on the performance of an IEEE 802.16j mobile multi-hop relay (MMR) WiMAX network. *Digital Commun. Netw.* **4**(1), 58–68 (2018)
11. S. Kumarapandian, M.J. Sibley, Complexity analysis of double-threshold based relay selection in D2D cooperative network. *J. Wirel. Netw. Commun.* **8**(1), 1–6 (2018)
12. H. Adam, E. Yannmaz, C. Bettstetter, Medium access with adaptive relay selection in cooperative wireless networks. *IEEE Trans. Mob. Comput.* **13**(9), 2042–2057 (2014)
13. M. Qian, C. Liu, Y. Zou, Cooperative beamforming for physical-layer security in power-constrained wireless sensor networks with partial relay selection. *Int. J. Distrib. Sens. Netw.* **12**(3), 9740750 (2016)
14. K. Shim, N.T. Do, B. An, Performance analysis of physical layer security of opportunistic scheduling in multiuser multirelay cooperative networks. *Sensors* **17**(2), 377 (2017)
15. G. Amarasinghe et al., Output threshold multiple relay selection scheme for cooperative wireless networks. *IEEE Trans. Veh. Technol.* **59**, 3091–3097 (2010)
16. Y. Zou, X. Wang, W. Shen, Optimal relay selection for physical-layer security in cooperative wireless networks. *IEEE J. Sel. Areas Commun.* **31**(10), 2099–2111 (2013)
17. Kun Xie, Jian-Nong Cao, Ji-Gang Wen, Optimal relay assignment and power allocation for cooperative communications. *J. Comput. Sci. Technol.* **28**(2), 343–356 (2013)

Analysis of Performance of FSO Link During the Months of Monsoon in Delhi, India



Sanmukh Kaur, Syed Zafar Ali Raza, Jaideep Khanna and Anuranjana

Abstract The FSO communication links use an open free space for transmission and thus must be designed to withstand the atmospheric challenges which affect the capacity of the system. Rain is one of the foreign elements which can deteriorate the performance of the link and cause huge effect on the reception of the signal. In this paper, we present a comprehensive survey of attenuation due to rain conditions in FSO link in the months of monsoon in Delhi region. The rain attenuation used for simulation of the system has been calculated by using Marshal and Palmer rain distribution model for four specific months of the rainfall, i.e., June to September. Q-factor of the received signal has been analyzed by varying the transmission wavelength, data rate, and range of the FSO system. The simulation results show that for error-free transmission of data during the months of rainfall, the transmission signal wavelength in the longer wavelength region at a wavelength around 1550 nm is recommended to be used for a maximum data rate of 2.5 Gb/s with a transmission range of 4 km.

Keywords Rain attenuation · FSO link · Q-factor · Data rate

S. Kaur · S. Z. Ali Raza · J. Khanna · Anuranjana (✉)

Amity School of Engineering and Technology, Amity University, Noida, India
e-mail: aranjana@amity.edu

S. Kaur
e-mail: sanmukhkaur@gmail.com

S. Z. Ali Raza
e-mail: zafar.raza1214@gmail.com

J. Khanna
e-mail: jai.khanna1996@gmail.com

1 Introduction

Delhi is the capital of India, with a huge population of 11 million in 2011 standing after Mumbai. With an area of 1484 km^2 , it has seen a rapid growth up to 26 million in 2016 [1]. There is a relentless increase in demand for bandwidth in metro cities. So here we see a huge requirement of telecom technologies. One of the best solutions that came into being was the use of optic fiber in offices, homes, and even restaurants and malls. But still we are facing huge problems while installing and maintaining the cables. Use of radio frequency (RF) technology does offer a longer communication range but again RF based network requires a great capital investment to acquire spectrum license [2, 3]. As a result of increasing data and internet supply, congestion occurs that raises a need to switch from RF technology to free space optical (FSO) technology [4, 5].

FSO refers to free space optical communication technology that uses light propagating in free space to wirelessly transmit data for telecommunication. The light is generated in the form of a narrow beam, by a laser source for long-distance or by LED for short-distance transmission and is propagated through a channel which is air or vacuum in the wavelength window of 750–1600 nm [6]. It is one of the most viable alternatives as it gives an optimal solution in terms of bandwidth scalability, speed of installation, reinstallation, portability, license-free spectrum, and cost-effectiveness. The extremely narrow bandwidth of the transmission and absence of fresnel zones, side lobes or back lobes, makes the interception hard and thus keeps the transmission secure [7, 8].

FSO uses an open free space for transmission that is prone to its own type of disturbances. The networks must be designed to withstand the atmospheric challenges which affect the capacity of the system. With the mentioned advantages, the main challenges of this type of communication system remain as attenuation or distortion of the signal under poor weather conditions [9].

Presence of foreign elements like rain, fog, and haze or any physical obstruction can deteriorate the performance of the link and can cause a huge effect on the reception of the signal. The diameter of waterdrops is larger and hence the chances of data loss and attenuation tend to increase during the months of rainfall. In this paper, the impact of rain on FSO communication link has been studied using Marshal and Palmer rain attenuation model. We have collected the rainfall data for the years 2012–2016 during the months of monsoon, i.e., June–September in Delhi region and analyzed the extent up to which the performance of the link may be affected considering different values of wavelengths, data rates, transmission ranges, and modulation formats.

2 System Layout

The proposed FSO system has been simulated for performance characterization using optisystem-15. Optisystem-15 is an innovative optical simulation package used for designing, testing, and optimization of virtually any type of optical system in the

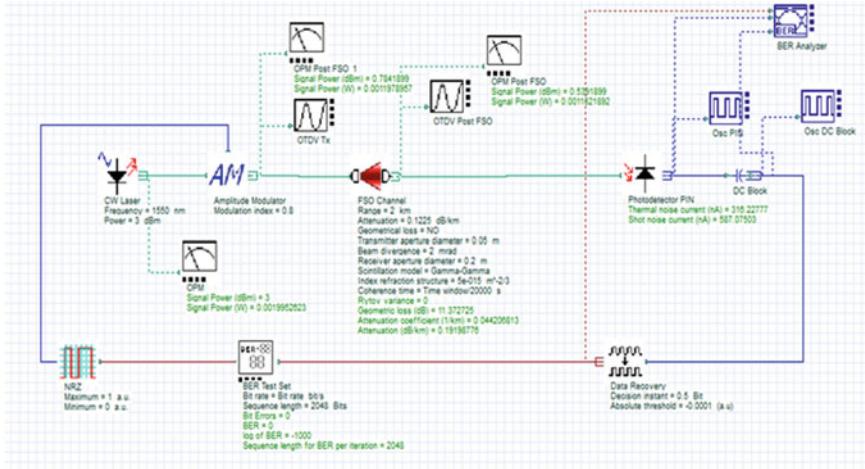


Fig. 1 System model analyzing the effect of rain attenuation

physical layer of spectrum of optical networks. Its use can minimize the time required and decreases the overall cost related to the design of an optical system. The FSO system basic design model is shown in Fig. 1. In FSO system, the transceivers used for communication at both ends should be in line of sight for successful transmission of signal. In the proposed design, the communication link has three different sections including Transmitter, propagation channel, and Receiver. The different subsystems are non-return-to-zero (NRZ) pulse generator, continuous wave (CW) laser, amplitude modulator, FSO transmission link, demodulator, and bit error rate (BER) analyzer. The NRZ pulse generator generates NRZ pulses coded by an input digital data signal. The CW laser generates a light output at a wavelength of 1550 nm. The modulated light output signal from amplitude modulator is applied to the FSO channel consisting of telescopes at both the ends. The pin photo detector is an integral part of the receiver and performs the regeneration of an electrical signal corresponding to original bit sequence transmitted. BER analyzer has been used to evaluate the BER and Q-factor of the received optical signal at the other end.

3 Degradation of Performance of FSO Link Due to Effect of Rain

As per the dimensions of the raindrop, there are many distribution designs which may determine the Rain attenuation model [10]. We have analyzed using the most commonly used rain distribution model which is marshal and palmer rain attenuation model. Marshal and Palmer's distribution model prescribes their Laws and facts to estimate the specific rain attenuation using an empirical formula [11]. The specific attenuation formula for rain effect analysis is given as:

Table 1 Rainfall data of Delhi Region for months June–September

Year	Rain (mm)				Rain (mm) June–September
	June	July	August	September	
2012	7	113.4	222.6	65.8	408.8
2013	110.9	189.3	177.9	58.4	536.5
2014	27.1	111.1	80.0	71.2	289.4
2015	58.9	268.8	244.7	26.1	598.5
2016	59.7	312.0	103.1	48.0	522.8
2017	103.8	109.7	117.0	112.1	442.6

Table 2 Specific attenuation per year based on average rain rate (mm/h)

Year	Rain (mm) June–September	Rain rate (mm/h)	Attenuation (dB/km)
2012	408.8	0.1396	0.1055
2013	536.5	0.1832	0.1253
2014	289.4	0.0988	0.0849
2015	598.5	0.2044	0.1342
2016	522.8	0.1785	0.1233
2017	442.6	0.1512	0.1110

Average attenuation = 0.11 dB/km

$$\gamma \text{ (dB/km)} = k R^\alpha \quad (1)$$

where k and α are constants whose values depend upon wavelength, temperature, and raindrop size distribution and R is the rain rate in (mm/h). For spherical raindrops and operation at a wavelength of 1550 nm, k and α are given as 1.076 and 0.66, respectively [11].

As the statistics given by the meteorological department is monthwise in mm, average rain per year during the months of monsoon has been computed as depicted in Table 1. It has been further converted to the average rain rate, i.e., Rain/hour (mm/h) per year for the months of rainfall. The specific attenuation in dB/km per year has been calculated thereafter based on average rain rate as shown in Table 2.

4 Simulation Results

This section includes the observation of results and their discussion using simulation software Optisystem-15. The proposed FSO system has been optimized for improving the overall performance of the link in a region where rain attenuation affects the effectiveness of the link during the months of monsoon. The data rate and wavelength

Table 3 Default parameters

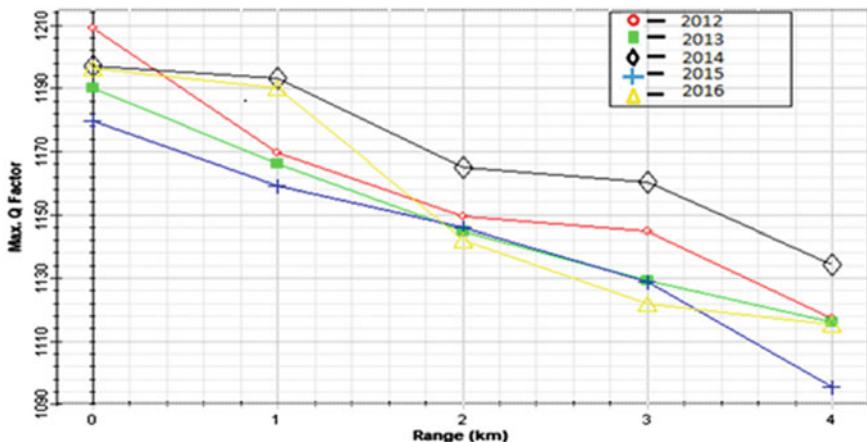
Parameter	Value
Data rate	10 Gbps
Optical Tx power	3 dBm
Wavelength	1550 nm
Modulation format	NRZ
Transmitter aperture diameter	0.05 m
Receiver aperture diameter	0.2 m
Range	2 km

used for analysis have been fixed at 10 Gb/s and 1550 nm, respectively. The other default parameters of the system are listed in Table 3.

We have analyzed the performance of the link by collecting the rainfall data for the years 2012–2017 during the months of monsoon, i.e., June–September in Delhi region [12]. For the calculated value of attenuation as shown in Table 2, Q-factor and received signal power have been analyzed first with respect to transmission wavelength and range for the years 2012–2016. The performance of the FSO system is further measured for different data rates and modulation formats considering average attenuation 0.11 dB/km during months of rainfall.

The graph in Fig. 2 shows the inverse relationship between Q-factor and range for years 2012–2016. The transmission range has been varied from zero to 4 km. The lowest value of Q-factor has been observed for the year 2015 with the highest attenuation of 0.13 dB/km.

In Fig. 3, Q-factor has been estimated for a wavelength range of 700–1600 nm covering three transmission windows of optical propagation during months of monsoon for the years 2012–2016. It has been observed that the higher value of Q-factor

**Fig. 2** Q-factor versus transmission range during the months of rainfall for the years 2012–2016

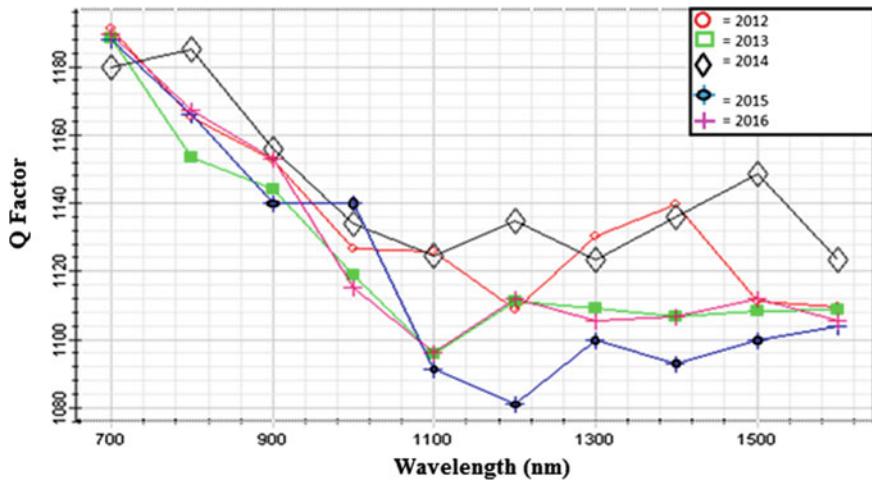


Fig. 3 Q-factor versus wavelength during the months of rainfall for the years 2012–2016

has been obtained for the first transmission window, i.e., 700–800 nm and almost similar values have been observed for the other two at 1330 and 1550 nm, respectively. FSO links operate around the wavelengths of 850 and 1550 nm. 1550 nm (about 200 THz) is a preferred choice because of eye safety [13].

For the calculated values of attenuations per year (Table 2), the received signal powers have also been observed with respect to transmission range from 0 to 2 km.

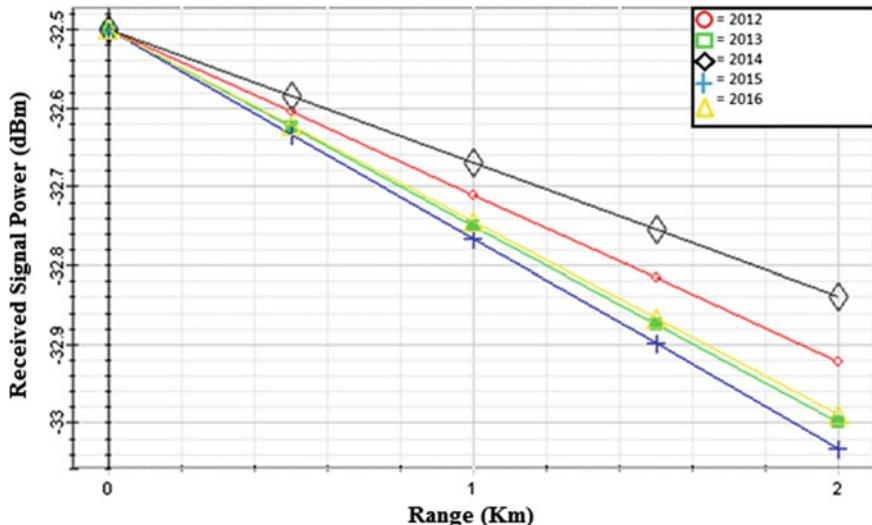


Fig. 4 Received signal power versus transmission range during the months of rainfall for the years 2012–2016

It can be observed in the graphs that received optical power has an inverse relation with respect to transmission range (Fig. 4).

The Plots of Figs. 5 and 6 have been obtained by considering the calculated value of average attenuation of 0.11 dB/km. Figure 5 depicts the plot of Q-factor of the received signal as a function of transmission range for varying data rates under average rain conditions. Here the Q-factor has been measured for a sequence of data rates ranging from 1 to 2.5 Gbit/s.

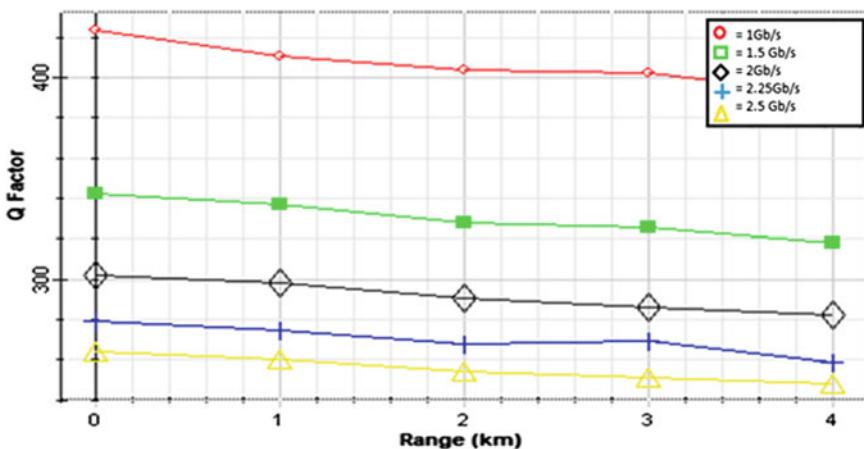


Fig. 5 Q-factor versus transmission range for different data rates

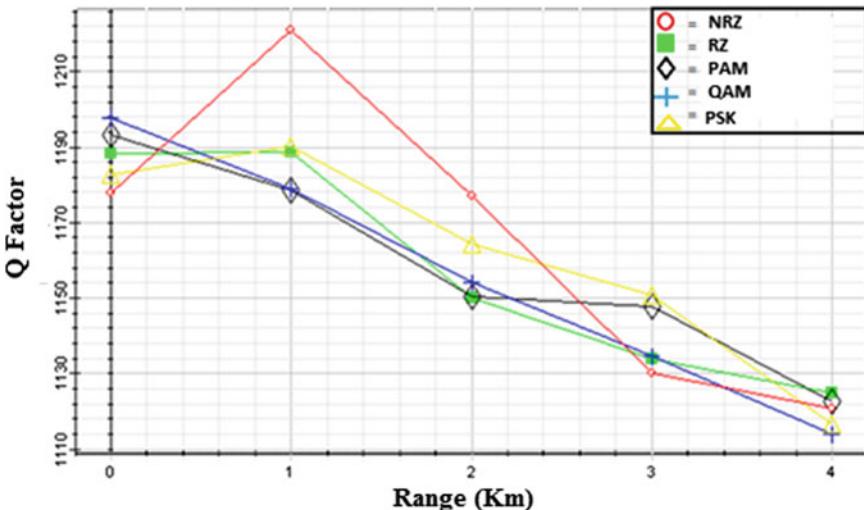


Fig. 6 Q-factor versus transmission range for different modulation formats

It can be observed from the figure that the quality of the received signal deteriorates with the increase in the data rate. An acceptable quality of the received signal has been obtained at a maximum transmission data rate of 2.5 Gbit/s up to transmission range of 4 km.

It is important to evaluate the performance of the link considering different modulation formats. We have considered five types of modulation schemes and observed the quality of the received signal as a function of deployment distance up to 4 km. It can be observed from the Fig. 6 that non-return-to-zero (NRZ) modulation scheme performs better than the phase shift keying (PSK) up to a link distance of 2.4 km. As the transmission distance is further increased, PSK performs better than the other modulation schemes.

5 Conclusion

In this paper, the impact of rain on FSO communication system has been studied using Marshal and Palmer rain distribution model. We have collected the rainfall data for the years 2012–2016 during the months of monsoon, i.e., June–September in Delhi region. Q-factor of the received signal has been analyzed by varying the transmission wavelength, data rate, and range of the FSO system. The simulation results show that for error-free transmission of data during the months of rainfall, the transmission signal wavelength in the longer wavelength region at a wavelength of 1550 nm is recommended to be used for a maximum transmission data rate of 2.5 Gb/s with a transmission range up to 4 km.

References

1. United Nations, Department of Economic and Social Affairs, Population Division (2016). The World's Cities in 2016—Data Booklet (ST/ESA/SER.A/392)
2. H. Kaushal, G. Kaddoum, Optical communication in space: challenges and mitigation techniques. *IEEE Comm. Surv. Tut.* **19**(1), 57–96 (2016)
3. S. Kaur, A. Kakati, Analysis of free space optics link performance considering the effect of different weather conditions and modulation formats for terrestrial communication. *J. Opt. Commun.* (2018). <https://doi.org/10.1515/joc-2018-0010>
4. S. Kaur, Analysis of inter-satellite free-space optical link performance considering different system parameters. *Opto-Electron. Rev.* **27**, 10–13 (2019)
5. F.U. Rashidi, S.K. Semakuwa, Performance analysis of free space optical communication under the effect of rain in Arusha region, Tanzania. *IJERT* **3**(9) (2014)
6. A. Kesarwani, A. Sharma, S. Kaur, M. Kaur, P.S. Vohra, Performance analysis of FSO link under different conditions of fog in Delhi, India, in *IEEE 2nd International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES-2018)*
7. N. Tabassum, N. Franklin, D. Arora, S. Kaur, Performance analysis of free space optics link for different cloud conditions, in *IEEE 4th International Conference on Computing, Communication and Automation (ICCCA 2018)*

8. S. Bloom, E. Korevaar, J. Schuster, H. Willebrand, Understanding the performance of free-space optics. *J. Opt. Netw. Opt. Soc. Am.* **2**(6), 178–200 (2003)
9. S. Chaudhary, A. Amphawan, The role and challenges of free-space optical systems. *J. Opt. Commun.* **35**(4), (2014)
10. J.S. Marshall, W.M. Palmer, The distribution of raindrops with size. *J. Meteorol.* **5**(4), 165–166 (1948)
11. Recommendation ITU-R P.838-3 Specific attenuation model for rain for use in prediction methods (Question ITU-R2013)
12. Website link, [http://hydro.imd.gov.in/hydrometweb/\(S\(1wex4oerw502fzygexmuu5jn\)\)/PRODUCTS/Publications/Rainfall%20Statistics%20of%20India%20-%202017/Rainfall%20Statistics%20of%20India%20-%202017.pdf](http://hydro.imd.gov.in/hydrometweb/(S(1wex4oerw502fzygexmuu5jn))/PRODUCTS/Publications/Rainfall%20Statistics%20of%20India%20-%202017/Rainfall%20Statistics%20of%20India%20-%202017.pdf)
13. O. Bader, C. Lui, Laser safety and the eye: hidden hazards and practical pearls, Technical Report: American Academy of Dermatology, Lion Laser Skin Center, Vancouver and University of British Columbia, Vancouver, B.C. (1996)

Pectoral Muscle and Breast Density Segmentation Using Modified Region Growing and K-Means Clustering Algorithm



Jyoti Dabass

Abstract Breast Cancer is the most habitually detected neoplasm amid women in India and it is one of the principal reasons for cancer decreases in females. In order to visualize the breast cancer, radiologists prefer to use mammogram. It consists of many artifacts, which negatively influences the detection of breast cancer. Presence of pectoral muscles makes abnormality detection a cumbersome task. The recognition of glandular tissue in mammograms is imperative in evaluating asymmetry between left and right breasts and in guesstimating the radiation risk connected with screening. Thus, the proposed technique focuses on breast part extraction, muscle part removal, enhancement of mammogram, and segmentation of mammogram images into regions conforming to different densities. The anticipated method has been verified on Mini-MIAS database mammogram images with ground truth offered by expert radiologists. The results show that the proposed technique is efficient in removing pectoral muscles and segmenting different mammographic densities.

Keywords Breast cancer · Image segmentation · K-means clustering · Mammograms · Mini-MIAS database · Pectoral muscles region growing

1 Introduction

Breast cancer is the utmost persistently analyzed tumor in the immense mass of the countries (154 of 185) incorporated in GLOBOCON 2018 and is also the primary reason of cancer demise in over 100 countries (15%). Wide-reaching, there are about 2.1 million recently diagnosed women breast malignancy cases in 2018, bookkeeping for almost one in four melanoma cases among female. In terms of death, breast cancer rates demonstrate less inconsistency with the utmost mortality anticipated in Melanesia, where Fiji has uppermost mortality rates worldwide [1]. It is also projected that by 2025, there will be 19.3 million new cancer cases [2, 3]. Furthermore, in developing countries like India, the mortality rate is high owing to patients' unawareness

J. Dabass (✉)

EECE Department, The Northcap University, Gurugram, India

e-mail: jyotidabas91@gmail.com

to the ailment symptoms, intense population, and looking for therapeutic discussion either when it's extremely critical or too late. Also, early detection and diagnosis of a breast tumor are difficult in rural areas due to less medical experts and specialists which augment the mortality rate. To condense the transience frequency and to boost the cure chances, medical support systems using medical data and information technology can be a great solution as it can assist in early detection of abnormalities by impersonating the doctor's reasoning and concluding symptoms. The accuracy and efficiency of the medical support system are increased by providing the exact region of interest. Extracting region of interest is a challenging task in preprocessing because the presence of pectoral muscles influences the detection of abnormality.

Also, breast tissue density (amount of glandular and fibrous tissue that looks on a woman's mammogram) disturbs the radiologist's aptitude to spot breast cancer. A cancerous lump can display up as white on a mammogram. Calcifications which may sometimes be connected with breast cancer or DCIS (ductal carcinoma in situ) also seem white on a mammogram. If the patient's breast tissue is dense, it may be like attempting to treasure a snowball in the blizzard. High tissue density can make it tougher to see certain vicissitudes on a mammogram that might eventually be cancer.

The BIRADS classification system is one of the most commonly used systems for breast density classification and it is extensively acknowledged in peril management and eminence assertion in mammography by the American college of radiology. BIRAD is classified into different stages. BIRAD I and BIRAD II are beginning of benign phases as breast cysts and breast Lipomas. BIRAD III designates probable benign stage and BIRAD IV characterizes skeptical anomaly and directing to malignant stages. The BIRAD classification demonstrates the prefigure of an upsurge in breast density in each phase in the MIAS database. The database has three classes of breast density: fatty, fatty glandular, and dense glandular and is established on BIRAD classification.

In order to assist the radiologists in early and accurate identification of breast cancer, the proposed method focused on breast part extraction, muscle part removal, enhancement of mammogram, and segmentation of mammogram into distinct densities. The layout of the paper is as trails. Section 1 presents the topic followed by the propositioned technique in Sect. 2. Section 3 delivers the outcomes and discussion while Sect. 4 ends the topic.

2 Related Work

2.1 *Image Segmentation*

The image segmentation step involves identifying a region of interest which could be done automatically, semiautomatically, or manually. Although manual segmentation is accurate, it is more tedious and subjective [4]. Automatic segmentation is

Table 1 Techniques for pectoral muscle segmentation

S. no	Author and year	Techniques used
1	Pavan et al. 2018 [12]	Hough transform for edge detection and active contour for segmenting pectoral muscles
2	Shinde, Rao 2019 [13]	Region growing, thresholding, and k-means clustering
3	Yin et al. 2019 [14]	Iterative threshold, rough and fitting contour
4	Shen et al. 2018 [15]	Genetic process, morphological medley algorithm, and polynomial curve fitting

objective but error-prone especially when imaging artifacts and noise are encountered. Some of the commonly used automated segmentation techniques include active contour-based [5], level set-based [6], and region- and graph-based methods [7–9]. No segmentation standard currently exists. More recently, deep-learning schemes such as convolutional neural networks have been exercised for segmentation [10, 11].

2.2 *Pectoral Muscle Segment*

Pectoral muscle (PM) is a high-density tissue, with the same imaging features as fibro glandular tissues, which figures its impulsive exposure a thought-provoking job [12]. So, it is essential to segment the pectoral muscles for early and accurate detection of breast cancer. Table 1 reviews the techniques used for segmenting pectoral muscles.

2.3 *Breast Density Segmentation*

Breast cancer typically happens in the fibro glandular part of breast tissue which appears bright on mammograms and is designated as breast density. The breast density share comprises ducts, lobular elements, and fibrous connective tissue of the breast. Breast density is a significant aspect in the elucidation of mammograms. The share of fatty and fibro glandular tissue of the breast region is assessed by the radiologists in the analysis of mammographic images. The upshot is idiosyncratic and fluctuates from one radiologist to another [16]. Therefore, it is imperative to segment the breast density. Table 2 reviews the breast density segmentation techniques.

Table 2 Techniques for breast density segmentation

S. no	Author and year	Techniques used
1	Mohamed et al. 2018 [17]	Convolutional neural network (CNN)-based model
2	Salman, Ali 2019 [18]	Region growing, median filter, and k-means clustering
3	Kallenberg et al. 2016 [19]	Deep convolutional network, sparse autoencoder
4	Oliver et al. 2015 [20]	Supervised pixel-based classification and exploiting textural and morphological features
5	Pavan et al. 2016 [21]	Fuzzy c-means clustering

3 Proposed Technique

The proposed technique involves four stages namely removal of background, suppression of pectoral muscle, image enhancement, and segmentation of the image into different breast densities. These stages are discussed below in detail. For implementing the proposed technique, publicly available mini-MIAS database is used. In this, the original MIAS Database (digitized at 50-micron pixel edge) has been abridged to 200-micron pixel edge and pared/expanded so that every image is 1024×1024 pixels.

3.1 Removal of Background

For extracting breast profile by removing background, mammogram image is binarized with threshold value 0.1. The proposed technique is able to work well for threshold value up to 0.7. After binarizing, connected components are organized in descending order for extracting breast profile or the largest blob with pectoral muscles.

3.2 Suppression of Pectoral Muscle

The second stage was used to reduce the pectoral muscle partly by expending an improved region growing method. The seeded region growing is one of the image segmentation approaches, it works in two behaviors based on selected pixel locational value and other is an assortment of seed point. The seed point may be selected manually or adaptively. In the proposed method, seed point is nominated automatically by considering the orientation of the mammography. This approach regulates the neighboring pixels of the seed point and scrutinizes whether the next pixels should be added to the region or not. The process is recapitulated till the complete region of interest is extracted.

3.3 Image Enhancement

The third stage is to enhance the quality of the image using two-stage adaptive histogram equalization techniques. The segmented image is enhanced and then provided to k-means clustering for segmentation based on breast density. The result of k-means clustering is a mammogram fragmented into areas of dissimilar density/textture. The cataloging is completed in a supervised training manner, bestowing to the expert radiologist's ground truth.

3.4 K-Means Clustering

The fourth stage is k-means clustering which practices iterative modification to provide an ultimate outcome. The algorithm inputs are the number of cluster k and the dataset. The dataset is an assortment of features for each data points. The algorithm twitches with original guesstimates for the k centroids which can be either arbitrarily produced or erratically nominated from the dataset. The process then recapitulates between two stages.

- *Data assignment step:* Each centroid outlines one of the clusters. In the step, each data point is dispensed to its nearest centroid based on the squared Euclidean distance.
- *Centroid update step:* In this step, the centroids are totaled. This is finished by captivating the mean of all data points dispensed to that centroid's cluster. The algorithm reiterates between steps one and two until a stopping criterion is met.
- *Choosing k :* To treasure the number of clusters in the data, the user is required to run the k-means clustering process for a range of k values and relate the outcomes.

3.4.1 Algorithm Steps for Image Segmentation Using K-Mean Clustering

- Let \mathcal{I} be an input image, \mathcal{ML} be a member label. $i = 1$ to M , $j = 1$ to N , D_{ijk} be distance between i th row, j th column, and k th cluster centroid, T be the maximum number of iteration, \mathcal{CS} be cluster centers $k = 1$ to k , and \mathcal{D} be matrix distance between input image \mathcal{I} and cluster centroids. ML_{ij} be label of i th row and j th column pixel of the input image.
- Set $T = 50$, $eps = 1e - 5$, set $nc = 4$, $CS = \text{random_initialize}$
- While $t < T$ and $cms > eps$ then $\mathcal{D} = \text{find_distance}(\mathcal{I}, CS)$, $\mathcal{ML} = \text{cluster_labelling}(\mathcal{D})$, $\mathcal{CS} = \text{update_cluster_centroid}(\mathcal{ML})$, $cmx = \max(\text{abs}(pCS - CS))$, $t = t + 1$, $pCS = CS.\text{end}$. While $nc = \text{no of the cluster}$ and pCS be previous iteration cluster centroids.

4 Results and Discussions

The proposed technique is experimented on 322 mammogram images of the mini-MIAS database that covers all types of images such as fatty, glandular, and dense (Fig. 1).

The background artifacts are removed by binarization with threshold 0.1 and all the connected components are organized into largest to smallest in size to extract the largest blob. Then the blob is multiplied with the original image to get the original breast profile. The region growing method practiced lessening the pectoral muscle part. The proposed method helps to select the seed point automatically to remove the pectoral muscles using an adapted region growing method. The conventional assortment of seed point is modified grounded on the alignment of the image. The mini-MIAS dataset consists of either left-oriented or right-oriented images [22]. Hence, the seed point is either left topmost or right topmost first nonzero pixel. The

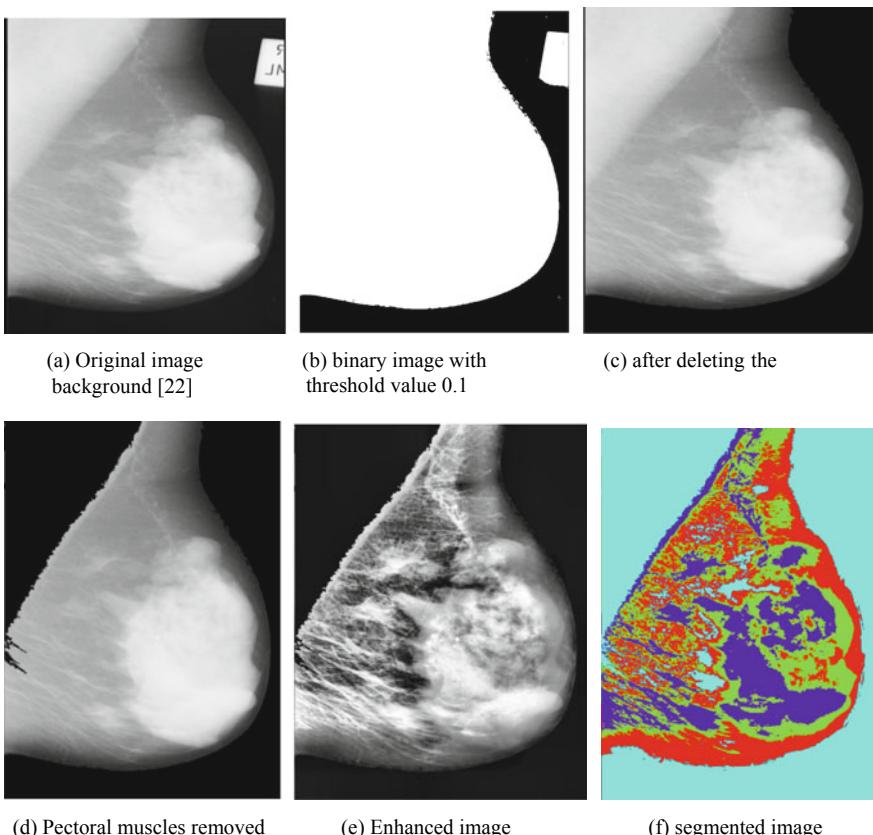


Fig. 1 Mammogram images obtained after pectoral muscle removal, enhancement, and segmentation

Table 3 Automatic region segmentation accuracy results

Segmentation accuracy	Dense	Fatty	Breast edge
98.75	95.8%	94.8%	

orientation of the image is found by dividing the image into half and counting the nonzero pixels; if left-oriented, left part consist of more pixels, else right part consists of more pixels. The goal of the anticipated segmentation process is to see if k-means clustering could disperse different densities (as described by Wolfe for the dissimilar density patterns) in the breast (adipose, glandular, etc.) according to what the radiologist professed as diverse density region while viewing a mammogram. The segmentation fallouts validate functionally conceivable breast density segmentation. According to radiologist valuation, the segmentation outcomes into sections with the breast edge, purely adipose tissue, adipose tissue with some ostensible erections (curvilinear structures or strands of fibrous tissue), comparatively dense tissue with some outward structure that may also embrace more see-through areas (fibrotic stromal tissue, glandular tissue), dense tissue with seeming texture structure (not homogeneous), and highly dense tissue (homogeneous). Table 3 gives the automatic areas segmentation accuracy outcomes for a three-category classification.

The results exhibit a robust resemblance between sundry textures and the radiologists' discernment of the various density areas in the breast. In the segmentation images ensuing from the presented algorithm, it can be discerned that darker the color, the lower the consequent density of the classified texture. The images are gauged quantitatively and qualitatively (where the radiologist's segmentation/classification is accessible). The algorithm was appraised on 322 normal mammogram images taken from a mini-MIAS database. For the qualitative assessment of the segmentation, a highly proficient breast screening radiologist was requested to scale the segmentation in one of the four groupings: very satisfactory, satisfactory, good, or poor. 310 images out of 322 images were graded as very satisfactory, 8 were assayed as satisfactory, and 4 as good.

5 Conclusion

We have imparted a technique for fragmenting the breast into regions of numerous density after removing pectoral muscles using modified region growing and enhancing the contrast using two-stage adaptive histogram equalization technique. The proposed technique is found efficient in removing pectoral muscles and segmenting breast images according to the density. The outcomes of the system show familiar concord to radiologist's dissection and texture density explanation. The presented approach incapacitates intricacies due to image procurement and breast inconsistency accomplishing a good exemplification of texture and density in the breast thus, delivering an excellent menace for density jeopardy evaluation, disproportionateness exposure, and pairing.

References

1. F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer J. Clin. **68**(6), 394–424 (2018)
2. Cancer IAFRo, Latest world cancer statistics Global cancer burden rises to 14.1 million new cases in 2012: marked increase in breast cancers must be addressed. World Health Organization, 2013. 12
3. F. Bray et al., Global estimates of cancer prevalence for 27 sites in the adult population in 2008. Int. J. Cancer **132**(5), 1133–1145 (2018)
4. C. Parmar, E.R. Velazquez, R.J. Leijenaar, et al., Robust radiomics feature quantification using semiautomatic volumetric segmentation. PLoS ONE **9**, e102107 (2014)
5. Z. Liu, L. Zhang, H. Ren, J.Y. Kim, A robust region-based active contour model with point classification for ultrasound breast lesion segmentation. SPIE Digital Library website, www.spiedigitallibrary.org/conference-proceedings-of-spie/8670/86701P/A-robust-region-based-active-contour-model-with-point-classification/10.1117/12.2006164.short?sso=1. Published on February 28, 2013. Accessed 11 Nov 2018
6. K. Suzuki, M.L. Epstein, R. Kohlbrenner, et al., CT liver volumetry using geodesic active contour segmentation with a level-set algorithm. SPIE Digital Library website, www.spiedigitallibrary.org/conference-proceedings-of-spie/7624/76240R/CT-liver-volumetry-using-geodesic-active-contour-segmentation-with-a/10.1117/12.843950.short. Published March 9, 2010. Accessed 11 Nov 2018
7. J. Peng, P. Hu, F. Lu, Z. Peng, D. Kong, H. Zhang, 3D liver segmentation using multiple region appearances and graph cuts. Med. Phys. **42**, 6840–6852 (2015)
8. W. Wu, Z. Zhou, S. Wu, Y. Zhang, Automatic liver segmentation on volumetric CT images using super voxel-based graph cuts. Comput. Math. Methods Med. **2016**, 9093721 (2016)
9. C. Sun, S. Guo, H. Zhang, et al., Automatic segmentation of liver tumors from multiphase contrast-enhanced CT images based on FCNs. Artif. Intell. Med. **83**, 58–66 (2017)
10. A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks. Neural Information Processing Systems website, <http://www.papers.nips.cc/paper/4824-image-net-classification-with-deep-convolutional-neuralnetworks.pdf>. Published in 2012. Accessed 11 Nov 2018
11. S. Pereira, A. Pinto, V. Alves, C.A. Silva, Brain tumor segmentation using convolutional neural networks in MRI images. IEEE Trans. Med. Imaging **35**, 1240–1251 (2016)
12. A.L. Pavan, A. Vacant, A.F. Alves, A.P. Trindade, D.R. de Pina, Automatic identification and extraction of pectoral muscle in digital mammography, in *World Congress on Medical Physics and Biomedical Engineering 2018* (Springer, Singapore, 2019), pp. 151–154
13. V. Shinde, B.T. Rao, Novel approach to segment the pectoral muscle in the mammograms, in *Cognitive Informatics and Soft Computing* (Springer, Singapore, 2019), pp. 227–237
14. K. Yin, S. Yan, C. Song, B. Zheng, A robust method for segmenting pectoral muscle in mediolateral oblique (MLO) mammograms. Int. J. Comput. Assist. Radiol. Surg. **14**(2), 237–248 (2019)
15. R. Shen, K. Yan, F. Xiao, J. Chang, C. Jiang, K. Zhou, Automatic pectoral muscle region segmentation in mammograms using genetic algorithm and morphological selection. J. Digital Imag. 1–12 (2018)
16. N. Saidin, H.A.M. Sakim, U.K. Ngah, I.L. Shuaib, Segmentation of breast regions in mammogram based on density: a review. Int. J. Comput. Sci. Issues (IJCSI) **9**(4), 108 (2012)
17. A.A. Mohamed, W.A. Berg, H. Peng, Y. Luo, R.C. Jankowitz, S. Wu, A deep learning method for classifying mammographic breast density categories. Med. Phys. **45**(1), 314–321 (2018)
18. N.H. Salman, S.I.M. Ali, Mammograms Segmentation and extraction for breast cancer regions based on region growing. Baghdad College Econ. Sci. Univ. **57**, 448–460 (2019)
19. M. Kallenberg, K. Petersen, M. Nielsen, A.Y. Ng, P. Diao, C. Igel, M. Lillholm, Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. IEEE Trans. Med. Imaging **35**(5), 1322–1331 (2016)

20. A. Oliver, M. Tortajada, X. Lladó, J. Freixenet, S. Garneau, L. Tortajada, R. Martí, Breast density analysis using an automatic density segmentation algorithm. *J. Digit. Imaging* **28**(5), 604–612 (2015)
21. A.L.M. Pavan, M. de Oliveira, M. Alvarez, A.J.M. Sampaio, A.P. Trindade, S.B. Duarte, D.R. de Pina, Breast tissue segmentation by fuzzy C-means. *Phys. Med.* **32**, 336 (2016)
22. J. Suckling et al., The mammographic image analysis society digital mammogram database Excerpta Medica. Int. Congr. Ser. **1069**, 375–378 (1994)

Author Index

A

- Ahmad, Mahira, 107
Ahmad, Tanvir, 183
Ali Raza, Syed Zafar, 321
Angne, Hemali, 255
Anuranjana, 321
Archana, E., 39
Atkari, Aditya, 255

B

- Bala Sai Krishna, Nadella, 295
Banati, Hema, 159
Bhargava, C. P., 13

C

- Chatterjee, Madhumita, 21, 137
Chouhan, Sonali, 275, 295

D

- Dabass, Jyoti, 331
Dabeer, Sumaiya, 107
Dange, Smita, 137
Das, Indrani, 235
Das, Sanjoy, 235
Dhargalkar, Nishant, 255
Divya, S., 59

G

- Gandhi, Jay, 285

H

- Hasan Khan, Muneeb, 107

Hassan, Syed Imtiyaz, 217

K

- Kannimoola, Jinesh M., 39
Kaur, Harkamaldeep, 119
Kaur, Manbir, 119
Kaur, Sanmukh, 321
Khanna, Jaideep, 321
Kiran, Eranki L. N., 59
Kumarapandian, Shamganth, 305
Kumar, Gautam, 1
Kuruvila, Aby, 39

M

- Manitpornsut, Suparerk, 265
Marathe, Nilesh R., 245
Mehta, Gitanjali, 195
Mishra, Manuj, 13

N

- Narayankutty, Revathi, 39
Narayan, Rohini, 195
Narmawala, Zunnun, 285
Negi, Anil Kumar, 217

P

- Pongdamrong, Prapas, 265

R

- Rajeev, Akshay, 39
Rao, Madhu Sudana, 59

S

- Saini, Hemraj, 1
Sarosh Umar, Mohammad, 107
Sharma, Ankita, 159
Shaw, Rabindra Nath, 235
Shinde, Subhash K., 245
Sibley, Martin James, 305
Siva Ganga Prasad, Muktyala, 69, 81
Srichareon, Benyatip, 265
Srikanth, Nandoori, 69, 81
Sufyan Beg, M. M., 183

T

- Taye, Tinamoni, 275

V

- Vemulapati, Pujitha, 59
Venkatesh, Bhukya, 295
Vyawahare, Madhura, 21

W

- Wazir, Samar, 183

Y

- Yadav, Pradeep, 13