RESEARCH JOURNAL OF FISHERIES AND HYDROBIOLOGY

2016. 11(4): 13-18
ISSN: 1816-9112
Journal home page: http://www.aensiweb.com/JASA/

Survey on map reduce based apriori algorithms in medical field for the prediction of diabetes mellitus

[1]Muni Kumar N and [2]Manjula R

[1,2]School of Computer Science & Engineering, VIT University, India.

Address For Correspondence:
Muni Kumar N, School of Computer Science & Engineering, VIT University, India.
Email: muni.kumarn2014@vit.ac.in, Mobile:+91-9000 33 0024

Received 22 January 2016; accepted 2 March 2016; published 25 March 2016

ABSTRACT

Now-a-days, Health care has become very expensive and greatest concern for everyone. With the changes in lifestyle, food habits and reduction in physical activities, most of the people are suffering with the attack of chronic diseases. Among several chronic diseases, Diabetes Mellitus was identified as one of the most common chronic disease being suffered by the people of all ages. Particularly, in India, the prevalence of diabetes mellitus is very high in number and the numbers are increasing at an alarming rate. In general, most of the people notice the diabetes at the advanced stages which creates complications thereafter. Always early detection of diseases during pre-diabetic stage is highly recommended. Further, with the usage of Electronic Health Records and Electronic Medical Records by healthcare practitioners, it is possible to detect the disease during the initial stages. But, the only problem is the storage and maintenance of digital health records which is growing at an exponential rate resulting the Big Data. The processing and analysis of big data requires distributed environment known as Hadoop. Our study is to predict the diabetes mellitus and its risk factors and also help the practitioners in making strategic decisions using data mining techniques in the Hadoop environment. As part of survey, in this paper, we report and describe various MapReduce based Apriori algorithms and their performance measures.

Key words: Apriori Algorithm for Hadooop, Big Data in Health care, Diabetes Mellitus, Hadoop, Map Reduce.

INTRODUCTION

Diabetes Mellitus [1] is a group of metabolic diseases characterized by elevated blood glucose levels resulting from defects in insulin secretion. Diabetes Mellitus is classified into two types, Type1 Diabetes Mellitus and Type2 Diabetes Mellitus. Type1 diabetes mellitus [1] (T1DM) is characterized by the null production of insulin by pancreas and requires daily insulin injections. Type 1 diabetes is usually diagnosed during childhood or early adolescence. Type 2 Diabetes is caused by insulin resistance in the liver, increased glucose production and over production of fatty acids which results in failure to produce sufficient insulin. The prevalence of diabetes [3] for all age-groups worldwide was estimated to be 2.8% in 2000 and 4.4% in 2030. Further, the total number of people with diabetes is projected to rise from 171 million in 2000 to 366 million in 2030. As per Diabetes Atlas [4], between 2010 and 2030, there will be a 69% increase in number of adults with diabetes in developing countries and a 20% increase in developed countries. Hence, we have focused our study towards the early detection of diabetes mellitus, as most of the adults would be suffering with diabetes which results in the reduction of human resources abilities. In most of the countries, health care costs account for a good percentage of its economy[5]. Health care industry is very critical and vast, unfortunately it is highly inefficient. Till-today, common people couldn't ripe the yield of cutting edge technology in the health care domain. Now it's time to bring reforms into the health care sector to improve the quality and care. This paper is organized as follows: Section II presents the Literature Survey, Section III focuses on the Big Data, Section IV

To Cite This Article: Muni Kumar N and Manjula R., Survey on map reduce based apriori algorithms in medical field for the prediction of diabetes mellitus. **Research Journal Of Fisheries And Hydrobiology**, 11(4): 13-18, 2016

describes the Hadoop Environment, Section V presents Data Mining, Section VI discusses about Apriori Algorithms and Section VII concludes the work.

II. Literature survey:

Cloud computing technologies and services over the internet bring the new improved health care solution. The proposed solution is replacement of the traditional health care strategy with the health care clouds. A health care cloud is the interconnection of extensive number of computers and servers especially dedicated to meet the needs to the health care industry. In the proposed solution, all the registered users who can be a doctor or a patient can access the services through the internet connection. The cloud services enable the registered user to access both the hardware and software managed by the third parties at remote locations. The advantages of developing the health care solution through cloud technologies include, Low cost computing services, Improved performance, Low cost of IT Infrastructure, Less Maintenance Issues, Low software cost, Universal access, Effective Collaboration, Improved Compatibility, Forecasting and Instant Software Update. Here, the main challenges would be user acceptance, proper bandwidth, infrastructure, security, maintenance and data analysis. Ahmed E. Youssef [7] has proposed a framework that paves a way for the generation of low cost, more efficient secure health care systems based on the big data analysis in mobile cloud computing environment. In this framework, EMRs and EHRs were integrated and made available for exchange by the doctors and patients. The important components of the framework include Cloud, EHR, Security model, Big data analysis, Google big query and Map Reduce. As part of our survey, we report that, this framework provides a high level of integration, interoperability and sharing of EHRs among health care providers, patients and practitioners. The implementation of said framework extends the health care access to the people living in any corner of the globe. The patients and the doctors have access to the framework such that, the patient can send a request to the doctor regarding his health and get the comments of the doctors with in short time. As the user base increases, the real problem arises with the storage and accessing of huge volumes of data. In real time , the data grows in huge volume, with high velocity in different varieties both structured and un-structured making the data "Big Data". Therefore, there is a need for the storage of huge volumes of data and further processing the data to gain insights and predict the health behavior of the patient. so, we investigate the various ways of analyzing the huge volumes of big data using distributed Hadoop environment. Big data analytics and Hadoop environment for storage and processing of huge volumes of data results in the reformation of the health care sector. Peter Augustine [8] has explained how the data flowing from various health monitoring devices become big data. Though big data analytics and Hadoop contribute major role in the storage and processing of huge volumes of health big data through distributed platform and Map Reduce techniques, there is a need for the data mining algorithms for the prediction and classification of data patterns. The data mining algorithms are to be employed to the processed data for the prediction of diabetes mellitus. Here, our survey will report the various algorithms which can be implemented on Hadoop distributed environment.

III. Big data:

Big data refers to the huge volume of data generated with high velocity. Generally, the social network data from the micro-blogging sites like Twitter, LinkedIn and social media platforms like Face book is considered as big data. Also, the data generated by the traditional enterprises in the form of transactional data and machine/sensor generated data like call data records, smart meters, manufacturing sensors, trading systems, traffic data, air data etc. which keeps on increasing without the human intervention is termed as Big Data. The following are the four characteristics of big data.

3.1. Volume:

Data volume has been increasing exponentially. i.e., at present, more than 2.5 Exa bytes of data is being generated and stored every day by various organizations ranging from Terabytes to Exa bytes and Zetta bytes of data.

3.2. Velocity:

Big data grows rapidly generating huge volumes that needs to be stored, transmitted, and processed quickly. Velocity is the speed at which the data is being generated. The data with high velocity includes the streamed data from various smart devices into social media and also camera streamed data that stores the data in motion from huge number of closed circuit cameras.

3.3. Variety:

In Big data, the variety and heterogeneity of data sources and storage has increased, fuelled by the use of cloud, web and online computing. Variety makes big data really big. Big data comes from a great variety of sources and generally has three types: structured, semi-structured and unstructured.

3.4. Veracity:

Big Data Veracity refers to the biases, noise and abnormality in data. Veracity in data analysis is the biggest challenge when compared to other characteristics like volume and velocity.

IV. Hadoop Environment:

Hadoop is an open source framework which supports the processing of the large data sets in distributed computing environment and it is part of the Apache project from Apache Software foundation. Hadoop is very much cost-effective for handling massive complex and heterogeneous data sets than traditional approaches. Hadoop is also a software library that can detect and handle failures in the cluster of commodity hardware at the application layer. The Hadoop architecture includes two main components.

4.1. Hadoop Distributed File System (HDFS):

The Hadoop Distributed File System [9] works with the cluster of commodity hardware. HDFS can handle huge volumes of data, as the data will be partitioned and stored across various nodes in the cluster. HDFS is highly fault tolerant and designed to be deployed on low-cost hardware. HDFS follows the Master-Slave architecture that contains two elements namely Data Node and Name Node. The Data Node actually stores the data blocks of the file and the Name node contains the meta data with records of blocks and listing of data nodes in the cluster.

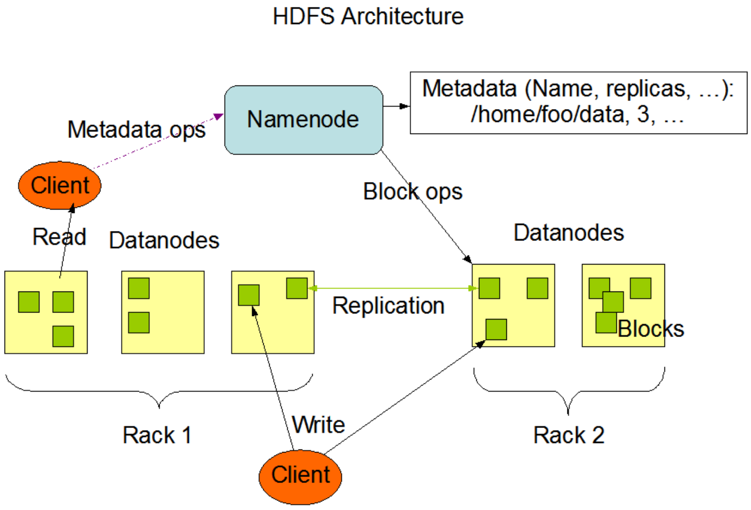

Fig. 1: HDFS Architecture

4.2. Map Reduce:

Map Reduce [10] is a software framework for easily writing applications to process vast amounts of data in parallel on large cluster of commodity hardware. A Map Reduce job usually splits the input dataset into independent chunks which are processed by the map tasks in parallel. The framework sorts the output of the maps, which are then input to the reduce tasks.

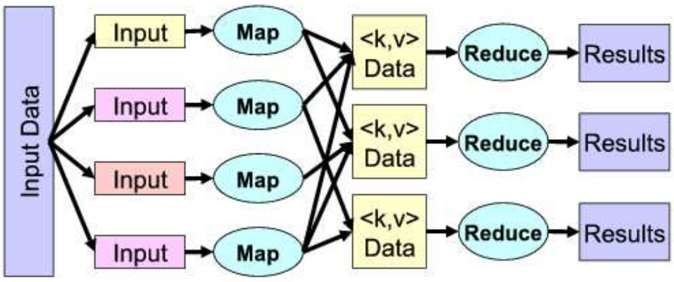

Fig. 2: Map Reduce

V. Data Mining:

Data Mining is the process of analyzing data from different perspectives and extracting the useful knowledge from huge volumes of data. It can also be defined as the extraction of knowledge from large databases/data sets.

In data mining, we have several data mining methods for classification and prediction of the data. In this paper our methodology is to categorize different data mining methods based on their functionality [11]. The data mining algorithms are classified majorly as follows.

- **Regression** is statistical methodology that is often used for numeric prediction.
- **Association** returns affinities of a set of records.
- **Classification** maps a data item into one of the predefined classes.
- **Clustering** identifies a finite set of categories to describe the data.
- Classification algorithms require that the classes be defined based on data attribute values. Pattern recognition is also a type of classification where an input pattern is classified into one of several classes based on similarity.

In our survey paper, we focus on the most popular Association Rule Mining Algorithm, Apriori and its variants on Hadoop platform. Apriori Algorithm will generate the association rules, finds frequent item sets satisfying minimum support threshold. We use these frequent item sets to produce association rules with minimum confidence value and that can predict the patients suffering with Diabetes Mellitus, also by changing the parameters in the dataset, we can even predict the pre-diabetes stages. Thus we go deep into the Apriori algorithm variants in next section.

VI. Apriori Algorithms:

Our survey has identified that, various data mining algorithms have been employed to assist the medical practitioner for the prediction any chronic disease during the early stages. Apriori algorithm is the most popular algorithm for mining frequent item sets. Based on the Apriori principle, any subset of frequent itemsets must also be frequent. In conventional databases, we can use the Apriori algorithm and predict the patient behavior. But, now we are aiming for the large data sets (Big Data Sets) and examine the behavior of Apriori principle when applied to Hadoop Map Reduce. When the dataset is large, the results would be more accurate and therefore the Hadoop based Apriori algorithm would produce better results than the conventional Apriori algorithm.

Various flavors of Apriori algorithm applied to Hadoop Map Reduce are:

Table 1: Apriori Algorithms

S.No	Author (s)	Name of the Algorithm
1	Ezhilvathani A & Raja. K	Parallel Apriori Algorithm
2	Wenqi Wang & Qiang Li	Improved Apriori Algorithm
3	Jongwook Woo	Apriori-Map/Reduce Algorithm
4	Wang L, Feng L, Zhang J & Liao P	FIMMR Algorithm
5	Liao J & Zhao Y	MRPrePost Algorithm
6	Sanjay R, Manohar K & Kashyap A	R-Apriori Algorithm
7	Yahya. O , Hegazy. O & Ezat. E	MRApriori Algorithm

i) Ezhilvathani has proposed the Parallel Apriori algorithm for frequent item set generation on Hadoop using Map Reduce programming model. But, the implementation part left to the future enhancement. Further, he discussed the implementation of Single node Hadoop cluster by considering a word count example. As part of our work, we will input diabetes dataset to this algorithm and generate the association rules which can detect the dependencies for the cause of Diabetes Mellitus.

ii) Wenqi have designed an improved Apriori algorithm for frequent item set generation suitable for Map/Reduce programming model and demonstrated that the new algorithm would generate better frequent item sets in association rule mining. This algorithm can predict the risk of diabetes among people of different age groups by finding the better frequent item-sets.

iii) Jongwook W proposed Apriori-Map/Reduce algorithm and illustrated the time complexity theoretically and proved the new algorithm gains much higher performance. The implementation was left as the future work. As this algorithm was theoretically proven, the decision of its usage can be decided after the implementation of algorithm and performance analysis.

iv) Wang proposed a frequent item set mining algorithm FIMMR for the big data environment. Conducted experiment on a platform of 27 nodes. Concluded that the time complexity of FIMMR algorithms is better than PFP and SPC algorithms. This algorithm can generate better association rules and thereby it produces the accurate measure of classification of diabetes patients.

v) Liao J proposed a parallel algorithm based on MapReduce. The experiment was conducted to compare the performance of PrePost, MRPrePost and PFP algorithms. Finally concluded, MRPrePost algorithm is suitable for large-scale data sets for mining association rules. As this algorithm is best suited for the large scale data sets, this can perform well with our diabetic big data sets with more number of attributes and huge population.

vi) Sanjay Rathee have proposed a Reduced-Apriori (R-Apriori), a parallel algorithm based on the Spark RDD Framework and evaluated the performance of R-Apriori and compared to Standard Apriori on Hadoop and concluded R-Apriori gives improved performance as the size of the dataset and the number of items increases. This algorithm makes use of the SPARK component of Hadoop framework and also algorithm eliminates the candidate generation steps, which can help the practitioners in making decisions in diabetes in quick time.

vii) Yahya have proposed a new algorithm based on Hadoop-MapReduce model called MapReduce Apriori Algorithm and conducted experiments with two existing algorithms and proved MRApriori is efficient and outperforms the other two algorithms. Here, this algorithm out performs the conventional Apriori methods applied to diabetic data sets.

Data Set:

The required big data datasets for the execution of the algorithm and prediction of Diabetes Mellitus is available in the following link.

http://archive.ics.uci.edu/ml/datasets/Diabetes

Further, the variety of datasets with huge number of attributes used for the research purpose may also be found in the above link.

The following table shows the performance of various Apriori algorithms applied for Hadoop Environment.

Table 2: Performance of Apriori Algorithms

S.No	Algorithm	Performance analysis
1	Parallel Apriori Algorithm	This algorithm was not implemented and modified the algorithms so as to use the Map Reduce method while candidate generation.
2	Improved Apriori Algorithm	Generates frequent item sets from massive data and make use of the full advantage of parallel processing.
3	Apriori-Map/Reduce Algorithm	Provides high performance computing depending on the number of map and reduce nodes.
4	FIMMR Algorithm	It discovers parallel local frequent item sets as candidates and then the candidates are filtered to achieve better performance over the conventional algorithms.
5	MRPrePost Algorithm	This algorithm is far better than Pre-Post and PFP. Also this is suitable for large scale datasets.
6	R-Apriori Algorithm	Here, the implementation was done on the Hadoop component, SPARK and this algorithm eliminates the candidate generation step and avoid costly comparisions.
7	MRApriori Algorithm	This algorithm outperforms one-phase and k-phase algorithms running on stand-alone mode of Map Reduce.

There are numerous opportunities in the health care domain, where the diagnostic models can be applied and various chronic diseases may be treated in better way. In our study, among the above 7 algorithms, we propose the following recommendations.

1. First Choice was to use the R-Apriori algorithm for the prediction of diabetes mellitus on the large diabetic dataset for built better support system to help diabetic patients for prediction and controlling of their blood glucose level.
2. Second choice is to use MRPrePost algorithm as it is highly suitable for large scale datasets(bigdata).

Conclusion:

In this paper, we discussed the need better health care and focused on the popular chronic disease, Diabetes Mellitus, cloud computing technology for the implementation and Bigdata for the storage and processing of the huge medical data. Further, in this paper we have compared various Map Reduce based Apriori algorithms to identify the best algorithm for the prediction of diabetes mellitus. This survey can help the young researchers to take the right path in designing the intelligent decision support system and help the practitioners in taking strategic decisions. The results of the algorithm can be used for the analysis of various clinical parameters, prediction of various diseases, forecasting tasks, medical knowledge extraction and patient management.

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my profound gratitude and deep regard to my guide, *Prof. R. Manjula, SCSE, VIT University,* for her exemplary guidance, valuable feedback and constant encouragement in completing this paper. Her valuable suggestions were of immense help in getting this work done. Her perceptive criticism kept me working to make this paper in a much better way. Working under her, was an extremely knowledgeable experience. Also, I would like to extend my sincere gratitude to my parents, my wife and my lovable daughter for their constant support and encouragement in completing this paper.

REFERENCES

1. ARENA, J.G., 2007. BEHAVIORAL MEDICINE CONSULTATION. *Handbook of Clinical Interviewing With Adults*, pp: 446.
2. Ruiz-Velázquez, E., A.Y. Alanis, R. Femat and G. Quiroz, 2011. Neural modeling of the blood glucose level for type 1 diabetes mellitus patients. In *Automation Science and Engineering (CASE), 2011 IEEE Conference on* pp: 696-701.
3. Wild, S., G. Roglic, A. Green, R. Sicree and H. King, 2004. Global prevalence of diabetes estimates for the year 2000 and projections for 2030. Diabetes care, 27(5): 1047-1053.
4. Shaw, J.E., R.A. Sicree and P.Z. Zimmet, 2010. Global estimates of the prevalence of diabetes for 2010 and 2030. Diabetes research and clinical practice, 87(1): 4-14.
5. Nambiar, R., R. Bhardwaj, A. Sethi and R. Vargheese, 2013. A look at challenges and opportunities of big data analytics in healthcare. In Big Data, 2013 IEEE International Conference on, pp: 17-22.
6. Chauhan, R. and A. Kumar, 2013. Cloud computing for improved healthcare: Techniques, potential and challenges. In E-Health and Bioengineering Conference (EHB), pp: 1-4.
7. Youssef, A.E., 2014. A framework for secure healthcare systems based on Big data analytics in mobile cloud computing environments. Int J Ambient Syst Appl, 2(2): 1-11.
8. Youssef, A.E., 2014. A framework for secure healthcare systems based on Big data analytics in mobile cloud computing environments. Int J Ambient Syst Appl, 2(2): 1-11.
9. Youssef, A.E., 2014. A framework for secure healthcare systems based on Big data analytics in mobile cloud computing environments. Int J Ambient Syst Appl, 2(2): 1-11.
10. Ayed, A.B., M.B. Halima and A.M. Alimi, 2015. MapReduce Based Text Detection in Big Data Natural Scene Videos. Procedia Computer Science, 53: 216-223.
11. Lakshmi, K.R. and S.P. Kumar, 2013. Utilization of Data Mining Techniques for Prediction of Diabetes Disease Survivability. International Journal of Scientific & Engineering Research, 4(6): 933-940.
12. Ezhilvathani, A. and K. Raja, 2013. Implementation of parallel apriori algorithm on hadoop cluster.
13. Woo, J., 2012. Apriori-Map/Reduce Algorithm. In *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA)* (p. 1). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).
14. Wang, L., L. Feng, J. Zhang and P. Liao, 2014. An efficient algorithm of frequent itemsets mining based on mapreduce. Journal of Information & Computational Science, 11(8): 2809-2816.
15. Wang, L., L. Feng, J. Zhang and P. Liao, 2014. An efficient algorithm of frequent itemsets mining based on mapreduce. Journal of Information & Computational Science, 11(8): 2809-2816.
16. Rathee, S., M. Kaul and A. Kashyap, 2015. R-Apriori: an efficient apriori based algorithm on spark. In Proceedings of the 8th Workshop on Ph. D. Workshop in Information and Knowledge Management., pp: 27-34.
17. Yahya, O., O. Hegazy and E. Ezat, 2012. An efficient implementation of Apriori algorithm based on Hadoop-Mapreduce model. In *Proc. of the*.