Analysis of Medical Domain Using CMARM: Confabulation Mapreduce Association Rule Mining Algorithm for Frequent and Rare Itemsets

Dr. Jyoti Gautam
HOD CSE and Associate Professor (CSE)
JSSATE
Noida, India

Neha Srivastava
Post Graduate of Computer Science & Engineering
JSSATE
Noida, India

***Abstract*—In Human Life span, disease is a major cause of illness and death in the modern society. There are various factors that are responsible for diseases like work environment, living and working conditions, agriculture and food production, housing, unemployment, individual life style etc. The early diagnosis of any disease that frequently and rarely occurs with the growing age can be helpful in curing the disease completely or to some extent. The long-term prognosis of patient records might be useful to find out the causes that are responsible for particular diseases. Therefore, human being can take early preventive measures to minimize the risk of diseases that may supervene with the growing age and hence increase the life expectancy chances. In this paper, a new CMARM: Confabulation-MapReduce based association rule mining algorithm is proposed for the analysis of medical data repository for both rare and frequent itemsets using an iterative MapReduce based framework inspired by cogency. Cogency is the probability of the assumed facts being true if the conclusion is true, means it is based on pairwise item conditional probability, so the proposed algorithm mine association rules by only one pass through the file. The proposed algorithm is also valuable for dealing with infrequent items due to its cogency inspired approach**.

Keywords—association rule mining; cogency; confabulation theory; medical data mining

I. INTRODUCTION

Data mining is one of the most vital and motivating area of research with the objective of finding meaningful information from huge data sets. In present scenario, Data mining is well accepted in healthcare field because there is a need of efficient analytical methodology for detecting unknown and valuable information in health data.

It also provides several benefits such as detection of the fraud in health insurance, availability of medical solution to the patients at lower cost, detection of causes of diseases and identification of medical treatment methods. It also helps the healthcare researchers for making efficient healthcare policies, constructing drug recommendation system, developing health profiles of individuals etc. Therefore, it can be concluded that Medical data mining has a great potential to discover the hidden relationship in the data sets of medical domain. This will help in understanding the prevailing situations in healthcare industry with respect to patients, their medical conditions and treatments. The most important thing in medical field is prognosis, i.e. an opinion, based on medical experience, of the likely course of a medical condition or the early diagnosis of any disease which helps in curing the disease at early stage and increase the life expectancy chances.

The data generated by the health organizations is very complex, heterogeneous in nature, and voluminous due to that it is difficult to analyze the data in order to make any important decision regarding patient health. So, there is a need to generate a robust, simple and computationally efficient tool for analyzing and extracting important information from such complex data sets. The development and understanding of such tools is the core business of data mining. Therefore, it is considered as an important area of study to explore the extent of association among such datasets. This is the reason that make association rule mining a crucial task.

Studies [11], [14] suggest that Confabulation association rule mining (CARM) approach, based on cognitive learning for generating association rules lead to generation of more interesting rules. CARM mine association rules by only one pass through the file as it is based on pairwise item conditional probability. Hence, it is more efficient for dealing with infrequent items. However, the confabulation based network is manually configured for selecting active items and interesting rules. So, Building such networks might require expert knowledge. Also, the value of parameters for finding items and for mining association rules must be re-configured when applied to different datasets. This limitation makes the Confabulation based approaches inflexible in handling different datasets. To address this limitation, an algorithm named CMARM: Confabulation MapReduce based association rule mining algorithm is proposed.

In the proposed algorithm, the concept of MapReduce framework of distributed computing is use for implementing the confabulation based association rule mining algorithm for both frequent and rare itemsets. MapReduce framework is regarded as the most successful computing platform for analyzing voluminous data as it adopts a data centric approach of distributed computing with the thought of "moving computation/processing closer to data". Also it provides higher level of abstraction which keeps many system level details hidden from the programmers and allows them to focus more on the problem explicit computational logic.

Hadoop is use here for implementing the proposed algorithm which is an open source implementation of MapReduce framework written in Java. Also, Java is use to write map and reduce function in the discovery and the mining phase. To improve the performance and execution time of MapReduce job, data is compress while writing them in HDFS. The main focus of this work is to make an automatic procedure that learns the structure of a confabulation network from the input datasets and mine the association rules for frequent and rare itemsets by strengthening the knowledge link between itemsets.

The remainder of this paper is organized as follows. Section II outline the related works in finding frequent and rare itemsets and different algorithm used for association rules mining. The proposed CMARM algorithm is introduced in section III and here, also discuss the performance evaluation which will be done by comparing the proposed algorithm with those of classical methods by using a graph in terms of memory usage, execution time and number of produced rules. Lastly, section IV concludes this paper.

II. Related Works

Commonly, Association rule mining (ARM) approaches use the concept of support-confidence to measure the potency of association rules. This concept may work splendidly for frequent itemsets mining but it is not necessarily suitable when rare itemsets are required. It can be explain as if the value of minimum support is considered high, then only few interesting rules are generated; and if it is considered low, then many uninteresting rules may be generated [11]. Even for low minimum support, some interesting rules may be lost if their corresponding items are rare. One more problem of this concept is its computational complexity as it required multiple scanning of the data sets for finding the support of itemsets.

In recent years, the main focus is given to the nature inspired algorithms for solving many today's challenging problems, including different perspective to ARM. For instance, kuo and shih proposed ant-based algorithm in 2007 and also show that the proposed method is able to provide more condensed rules than Apriori method and also, computational time is reduced. Later, in 2010, Suneetha and Krishnamoorti proposed an Organized Transaction Selection Approach, in which the rules are generated by selecting transactions according to the highest order first basis and hence avoiding unnecessary patterns generation. Therefore, the major advantage of this approach is the reduction of database scans and hence overcome relatively higher time complexity of Apriori algorithm.

To improve the quality of the generated rules, some multi-objective algorithms have been developed with more measures considered, other than only confidence factor or predictive accuracy. Ghosh and Nath, in 2004, introduces comprehensibility, interestingness, and confidence factor as three interestingness measures of rules with the objective to model the association rule mining problem as a multi-objective problem [2]. The proposed algorithm used these three measures to strengthen the quality of generated rules but sampling decreases the accuracy. In 2014, A. Gupta, R. Arora, R. Sikarwar and N. Saxena proposed another algorithm for web usage mining using improved frequent pattern tree algorithm [1], in which the system operates in three phases namely; Preprocessing module, Apriori or FP growth algorithm module- comprises: frequent itemset generation followed by rules derivation. The main drawback of Apriori algorithm is that the candidate set creation is costly, especially if a large number of patterns exist. The main drawback of FP-growth algorithm is lacks of good candidate generation method. Future research can combine FP-Tree with Apriori candidate generation method to solve the disadvantages of both Apriori and FP-growth [1].

Later in 2014, M.Parmar, M. Sutaria and M. Joshi proposed an approach for finding out frequent item set using comparison based technique, in which the author used a MINHEAP tree algorithm in place of FP-Growth tree to find most frequent item which is already sorted and present in root of heap tree [12]. In the same year, 2014, K. Mutakabbir, S.Mahin and M. Hasan, find the frequent pattern within a genetic sequence using unique pattern indexing and mapping techniques and showed that the proposed algorithm required multiple scanning of the database [6]. Kuo proposed a particle swarm optimization – based approach for association rule mining and showed that it is more efficient and has better runtime than genetic algorithm (GA) as in GA for each fitness evaluation needs to calculate support and confidence for corresponding rules, which is time consuming. In comparison to the population– based algorithm, neural network is an exemplar that promises time efficiency and high performance. Logically, this can be proving by observing human's remarkable ability to efficiently process voluminous data. These networks are based on the concept of human nervous system, which is considered as one of the most intelligent creature on earth. The neural network is nonlinear model that is easy to use and understand compared to statistical methods and also non-parametric model while most of statistical methods are parametric model that need higher background of statistic. So far, there are only a few researches that have applied neural network to ARM technique. Self-organizing map (SOM) is type of Artificial Neural Network (ANN), which is widely used in many data mining tasks, such as clustering and dimensionality reduction. The reduction in dimensionality allows people to visualize and interpret what would otherwise be, for all intents and purposes, scrawled data. Nohuddin discovered frequent trends in social network and use SOM to cluster these trends for better understanding of the nature of the trends. Furthermore, Yeh proposed a new approach called association reasoning neural network (ARNN) based on multilayer perceptron- network, where the number of hidden neurons act as the support threshold that control the generation of rules with low support value.

In 2014, A. Soltani and M. R. Akbarzadeh proposed Confabulation inspired association rule mining (CARM) approach [11], based on cogency inspired measure for generating rules. Cogency inspiration can lead to more intuitive rules. Also, cogency-related computations need pairwise item co-occurrences; hence findings rules requires only one pass through the file. Since, file access, particularly for large files can be significantly time consuming, therefore it can be concluded that the proposed algorithm is superior to the

Apriori algorithm due to one-time file access. In this paper, the parameter for selecting active items and for selecting interesting rules are predefined by the user. It would be desirable to set these parameters automatically using data set statistics. Moreover, the use of matrices in the implementation of CARM seems to basically make this a problem of statistical analysis.

Recently, in March 2015, Mansurul A.Bhuiyan and Mohammad Al Hasan proposed a frequent subgraph mining algorithm called FSM-H [10]. Here, author used an iterative MapReduce based framework for FSM-H and showed that it is complete as it returns all the frequent subgraphs for a given user-defined support, and also efficient as it applies all the optimizations that the latest FSM algorithms adopt. This algorithm is limited to the size of a graph which is equal to the number of edges it contains.

In June 2015, Qiuwen Chen, Qing Wu, Morgan Bishop, Richard Linderman and QiuQiu proposed an algorithm for self-structured confabulation network for fast anomaly detection and reasoning [14]. In this paper, the author proposed an automatic procedure that learns the structure of a confabulation network from the incoming dataset. The constructed model consists of well-defined nodes that capture both spatial and temporal relations among the features of the dataset. This work will be further extended by improving the workload distribution among the network nodes so that testing instance can be dynamically assigned to multiple heterogeneous devices.

III. METHODOLOGY

The proposed algorithm CMARM is designed as an iterative MapReduce process using Confabulation approach to find out the association rules for frequent and rare itemsets. Confabulation theory is an information processing model of human cognition introduced by Hecht-Nielsen [19]. It is based on the concept of cognitive learning. Cognitive learning concerned with the acquisition of problem-solving abilities, intelligence and conscious thought. It uses existing knowledge and generates new knowledge; therefore, lead us to the generation of more intuitive rules.

A. MapReduce Framework Implemenation

MapReduce is a linearly scalable programming model that enables distributed computation over voluminous data. The model provides two abstract functions: map and reduce each of which defines a mapping from one set of key-value pairs to another. The map function corresponds to the "map" operation in functional programming, whereas the reduce function corresponds to the "fold" operation in functional programming. The input to a MapReduce job starts as data stored on the underlying distributed file system. The Mapper is applied to every input key-value pair to generate an arbitrary number of intermediate key-value pairs. The Reducer is applied to all values associated with the same intermediate key to generate output key-value pairs. The files (input and output) of MapReduce are managed by a distributed file system. Hadoop is an open source implementation of MapReduce framework written in Java languages.

B. CMARM Algorithm

In the proposed CMARM algorithm, an additional parameter along with support and confidence threshold values is used to select active items and interesting rules in association rule mining. The value of this additional parameter is computed using cogency approach that leads us to formulation of more intuitive rules.

The proposed algorithm is divided into two main modules:

1) Frequent and Rare itemsets discovery,

2) Generate Strong Association Rule using confabulation theory and cogency measure.

In the first module, MapReduce framework is used to represent the knowledge links with weak strength using the threshold value of support count. This phase results in generation of frequent and rare itemsets such that whose support count is greater than the minimum specified support count value. In the second module, the strong association rules are generated based on confabulation theory, which states;

$$Fr = \{x \in I \mid supp(x) \geq S_0\}$$

It considers S_1= Fr where S_1 is 1-itemsets. After finding all frequent and rare n-itemsets, the algorithm generates all rules using their support, confidence and interestingness parameter for selecting active items and interesting rules that strengthen the knowledge link between the itemsets. Here, the value of interestingness parameter is estimated by mapper in terms of changes for both user-feature and item-feature pair. Then, the Reducer calculate the sum of changes and apply the calculated sums to each item rating by updating appropriate feature and use this output value of reduce function for next iteration.

C. Design Model

The design model of proposed CMARM algorithm is shown below:

1) Initially, parallel scan first divide the dataset information horizontally into 'P' node subsets and distribute it to 'Q' nodes supersets.

2) The various 'P' nodes are then processed again using mapping function.

3) Then once the method is completed, every node scans its own information sets then generates set of Candidate item set Fr.

4) Initially, the support count of every Candidate itemsets is about to one. This Candidate itemsets Fr is split into R partitions and sent to 'R' nodes with their support value count. Here, Min_Support is also defining which describes value of minimum support count. The Candidate itemsets Fr is discarded whose support count is less than Min_Support.

5) Once the algorithm has collected the set of all frequent and rare itemsets of all sizes that survived the support threshold. The next step is to extract strong associated rules from frequent and rare itemsets.

6) CMARM uses MapReduce framework to extract significant rules from all frequent and rare itemsets. If all frequent and rare itemsets can fit in computer memory and if the processing time is not that big then Hash table data

structure can be used to hold the data thrown from the map function. In this case, the key will be the left-part and the value will be set of (right-part: support) entries for frequent and rare itemsets. In the distributed implementation of this step, data are thrown to distributed file system and the Map-Reduce middleware is responsible to sort the entries, to fetch them, groups and sent to the reduce functions. Here, Mapper also calculates the changes for both user-feature and item-feature pair. These changes are used as a parameter for selecting active items and interesting rules from the frequent and rare itemsets. Then, Reducer calculates the sum of changes and applies sums to each item rating by updating appropriate feature and sends this to mapper function again for next iteration.

D. Implementation

The CMARM algorithm will implement using Hadoop which is open source of MapReduce framework written in Java language. The Map and Reduce function in the discovering and the mining phase of the Association Rule Mining will be written in Java Language. The choice to use Java language is motivated on several grounds. First, the construction of a runnable jar file is the easiest way to run a distributed program with Hadoop. Second, a Java implementation allows future integration with the Apache Mahout parallel machine learning library, a widely used package among machine learning end-users and researchers alike. To improve the performance and execution time of MapReduce job, the data is compress while writing them in HDFS. Also, global counter is use provided by Hadoop to track the stopping point of the iterative mining task.

Following are the Steps to implement proposed CMARM algorithm with Map Reduce framework:-

Step 1: Maps the input dataset to N partitions, where N = number of slave machines.

Step 2: Reduce phase would take the immediate key-value pairs emitted in the map phase.

Step 3: Send them altogether to the master node for further collecting the number for count per item.

3.1 To generate frequent and rare itemsets in form of key-value pair. This describes the number of occurrences of individual itemsets.

3.2 To generate the candidate sets from the source data file. It first discard those items that occur minimum than the support threshold value by looking up the global Hash map list and then recursively call mapping() function.

Step 4: Mapper calculates changes for both user-feature and item-feature pair, then using this value along with support and confidence strong association rule is generated.

Step 5: Reducer calculates the sum of changes and apply sums to each item rating by updating appropriate feature and use output of this reduce for next iteration.

E. Performance Evaluation

The relative performance of CMARM algorithm will analyze by comparing against two mining algorithm, one is CARM: Confabulation association rule mining algorithm and other is FIN algorithm, which is variant of FP tree algorithm. The Performance analysis will be shown by using a graph in terms of memory usage, execution time and number of produced rules. To verify the CMARM algorithm, both synthetic and real data sets are used, which is obtained from the health data repository.

IV. Conclusion

The foremost objective of this paper is to analysis of medical data repository to find out rare and frequent occurring diseases that may occur with growing age. So that, human being can take early preventive measures to minimize the risk of diseases that may supervene with the growing age and hence increase the life expectancy chances. Since, the data generated by medical domain is very vast and heterogeneous in nature; therefore, a new ARM algorithm is proposed called CMARM to mine the association rules. CMARM algorithm uses an iterative MapReduce based framework inspired by cogency. This algorithm uses the concept of cognitive learning for making clinical decisions which are often made based on doctor's intuition and experience rather than on the knowledge rich data hidden in the database. The proposed work can be further enhanced and expanded for the automation of long-term prognosis of diseases

References

[1] A. Gupta, R. Arora, R. Sikarwar, and N. Saxena, "Web usage mining using improved frequent pattern tree algorithm," IEEE conference on communication system and network technologies, pp. 573-578, 2014.

[2] A. Ghosh, and B. Nath, "Multi-objective rule mining using genetic algorithm," Information science, vol. 163, pp. 123-133, June 2004.

[3] D. Agnihotri, K. Verma, and P. Tripathi, "Pattern and cluster mining on text data," IEEE conference on communication system and network technologies, pp. 428-432, 2014.

[4] Faraz Rasheed, and Reda Alhaji, "A framework for periodic outlier pattern detection in time-series sequences," IEEE transactions on cybernetics, vol. 44, no.5, pp. 569-582, May 2014.

[5] G. Liu, H. Zhang, and L. Wong, " A Flexible approach to finding representative pattern sets," IEEE transaction on knowledge and data engineering , vol. 26 , no. 7, pp. 1562-1574, July 2014.

[6] K. Mutakabbir, S. Mahin, and M. Hasan, " Mining frequent pattern within a genetic sequence using unique pattern indexing and mapping techniques," IEEE conference on informatics electronics & Vision, 2014.

[7] K. Srinivas, B. Kavihta Rani, and Dr. A. Govrdhan, "Applications of data mining techniques in healthcare and prediction of heart attacks," International journal on computer science and engineering, vol. 02, no. 02, pp. 250-255, 2010.

[8] L. Cagliero, and P. Garza, "Infrequent weighted itemset mining using frequent pattern growth," IEEE transactions on knowledge and data engineering, vol. 26, no. 4, pp. 903-915, April 2014.

[9] L. Wang, David, Wai-Lok, and R. Cheng, "Efficient mining of frequent item sets on large uncertain databases," IEEE transactions on knowledge and data engineering, vol. 24, no. 24, pp. 2170-2183, December 2012.

[10] M. A. Bhuiyan and M. A. Hasan, "An iterative mapreduce based frequent subgraph mining algorithm," IEEE transactions on knowledge and data engineering, vol. 27, no. 3, pp. 608-620, March 2015.

[11] M. R Akbarzadeh-T, and A. Soltani, "Confabulation inspired association rule mining for rare and frequent itemsets," IEEE transactions on neural networks and learning systems," vol. 25, no. 11, pp. 2053-2064, November 2014.

[12] M. Parmar, M. Sutaria, and M. Joshi, "An approach for finding frequent itemset done by comparison based technique," International journal of computer science and mobile computing," vol.3, issue. 4, pp. 996-1001, 2014.

[13] M. Saravanan, and R. RJ, "Performance study of association rule mining algorithms for dyeing processing system," Innovative systems design and engineering, vol. 2, no. 5, pp. 34-43, 2011.

[14] Qiuwen Chen, Qing Wu†, Morgan Bishop†, R. Linderman† and Qinru Qiu, "Self-structured confabulation network for fast anomaly detection and reasoning," Air force research laboratory, Information Directorate, 2015.

[15] R. Agrawal, R. Srikant and others, "Fast algorithms for mining association rules," IEEE conference on informatics electronics & vision, vol. 1215, pp. 487-499, 1994.

[16] R. D. Canlas Jr., "Data mining in healthcare: Current applications and issues," International Journal on computer science and engineering," vol. 26, 2009.

[17] R. Jin, "An efficient implementation of apriori association web mining," Proc. workshop on high performance data web mining, April 2011.

[18] S. Bai and S. Bai, "The maximal frequent pattern mining of DNA sequence," IEEE conference on informatics electronics and vision, pp. 23-26, 2009.

[19] Srikant R, Agrawal R., "Mining sequential patterns: generalizations and performance improvements," Proceeding of 5th international conference on extending database technology (EDBT'96), pp. 3–17, Mar 1996.

[20] S. Q. Wang, Y. B. Yang, Y. Gao, G. P. Chen, and Y. Zhang, "Mapreduce based closed frequent itemset mining with efficient redundancy filtering," Proc. IEEE 12th International conference data mining workshops, pp. 49–453, 2012.

[21] Yang Xiang, Philip R.O. Payne, and Kun Huang, "Transactional database transformation and its application in prioritizing human disease genes," IEEE transactions on computational biology and bioinformatics, vol. 9, no. 1, pp. 294-304, January 2012.

[22] Z. Zhao, Da. Yan and W. Ng, "Mining Probabilistically Frequent Sequential Patterns in Large Uncertain Databases, IEEE Transactions on Knowledge and Data Engineering , vol. 26 , no. 5, pp.1171-1184, May 2014.