DISCOVERING FREQUENT ITEM SET USING CONFABULATION-INSPIRED ASSOCIATION RULE MINING

***Ms. T. Malathi, **Dr. T. Senthil Prakash, ***Mr. K. Arun**

**II year M. E. (CSE), Shree Venkateshwara Tech Engg College, Gobi*
***Professor &HOD, Shree Venkateshwara Hi-Tech Engg College, Gobi*
****II year M. E. (CSE), Shree Venkateshwara Hi-Tech Engg College, Gobi*

ABSTRACT

In recent years, the development of computer technologies, such as data storage and data base management systems, has enabled storage of huge amount of data. Data mining techniques are methods for obtaining useful knowledge from these large databases. One of the main tasks of data mining is association rule mining (ARM), which is used to find interesting rules from large amounts of data. A new confabulation-inspired association rule mining (CARM) algorithm is proposed using an interestingness measure inspired by cogency. Cogency is only computed based on pairwise item conditional probability, so the proposed algorithm mines association rules by only one pass through the file. The proposed algorithm is also more efficient for dealing with infrequent items due to its cogency-inspired approach. The problem of associative classification is used here for evaluating the proposed algorithm. This project evaluates CARM over data sets. Experiments show that the proposed algorithm is consistently faster due to its one time file access and consumes less memory space than the Conditional Frequent Patterns growth algorithm. In addition, statistical analysis reveals the superiority of the approach for classifying minority classes in unbalanced data sets.

Keywords- *Association Rule Mining, Cogency, Confabulation Theory, Frequent Patterns.*

I. INTRODUCTION

Data mining refers to extracting or "mining" knowledge from large amounts of data. Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set. These tools can include statistical models, mathematical algorithm and machine learning methods. Consequently, data mining consists of more than collection and managing

data, it also includes analysis and prediction. Classification technique is capable of processing a wider variety of data than regression and is growing in popularity.

The data mining functionalities are Characterization, Discrimination, Association analysis, Classification, Prediction, Clustering, Outlier analysis, Evolution and deviation analysis. The goal of data mining is to unearth relationships in data that may provide useful insights. Data mining tools can sweep through databases and identify previously hidden patterns in one step.

Market basket analysis relates to its use in retail sales. It has also been used to identify the purchase patterns of the Alpha consumer. Alpha consumers are people that play a key role in connecting with the concept behind a product, then adopting that product, and finally validating it for the rest of society.

Associative classification is a rule-based approach to classify data relying on association rule mining by discovering associations between a set of features and a class label. Support and confidence are the de-facto “interestingness measures" used for discovering relevant association rules[6]. The support confidence framework has also been used in most, if not all, associative classifiers. Although support and confidence are appropriate measures for building a strong model in many cases, they are still not the ideal measures and other measures could be better suited[7][8].

Rare association rule mining has received a great deal of attention in the recent past. In this research they use transaction clustering as a pre-processing mechanism to generate rare association rules. The basic concept underlying transaction clustering stems from the concept of large items as defined by traditional association rule mining algorithms. While addressing the problem of finding interesting predictive association rules in datasets with unbalanced class distributions [3], following disadvantages[4] occurs.

- A two phase association mining often can be found to be time and resource consuming in case of larger incremental datasets.
- Due to conversion of the real-life data into market-basket domain, information loss occurs.
- Single objective function (i.e. based on only frequency of occurrence) based rules generation often can be found to be non-interesting.

To address these issues, we need an effective and efficient association mining technique that can extract reduced set of interesting rules from real-life datasets without transforming it into the market basket domain[5].

II. RELATED WORK

An approach proposed by Koh & Pears (2008) to cluster transactions prior to mining for association rules [9] [10], shows that pre-processing the dataset by clustering will enable each cluster to express their own associations without interference or contamination from other sub groupings that have different patterns of relationships. The results show that the rare rules produced by each cluster are more informative than rules found from direct association rule mining on the un partitioned dataset[9] [10].

B.Nath et al. [1] presents an efficient incremental association mining technique which can discover non-redundant, reduced set of interesting rules from continuous valued dataset, without

converting into the market-basket domain. Yuan Cao et al. [2], proposed a novel method SOMKE, for kernel density estimation (KDE) over data streams based on sequences of self-organizing map (SOM). In many stream data mining applications, the traditional KDE methods are infeasible because of the high computational cost, processing time, and memory requirement. To reduce the time and space complexity, they proposed a SOM structure in this paper to obtain well-defined data clusters to estimate the underlying probability distributions of incoming data streams.

Setzkorn et al. [9] introduces a technique that uses Frequent Pattern Mining and Self Organizing Maps (SOMs) to identify, group and analyze trends in sequences of time stamped social networks so as to identify "interesting" trends. In this study, trends are defined in terms of a series of occurrence counts associated with frequent patterns that may be identified within social networks. Li.Gu et al. [10] presented an algorithm which finds association rules based on a set of new interestingness criteria. A statistically guided method of selecting appropriate features has also been developed. Initial results have shown that the proposed algorithm can find interesting patterns from data sets with unbalanced class distributions without performance loss.

Associative classification [11] is a rule-based approach to classify data relying on association rule mining by discovering associations between a set of features and a class label. Rare association rule mining has received a great deal of attention in the recent past. In this research [12], they use transaction clustering as a pre-processing mechanism to generate rare association rules. The basic concept underlying transaction clustering stems from the concept of large items as defined by traditional association rule mining algorithms. They make use of an approach proposed by Koh & Pears (2008) to cluster transactions prior to mining for association rules.

III. FREQUENT ITEMSET MINING USING ASSOCIATION RULE MINING

Market basket dataset mainly used in large of item are maintains. The dataset is frequently update at the time transaction, and also maintains frequent items. Suppose enter the new frequent items in transaction update the original dataset process is slow because can't already used in dataset. So large memory use and maintain large stock items, in this problem solve association mining rule used market basket dataset. Association rule mining is one of the major tasks of the Data Mining. Association rule mining finds the interesting or correlation relationship among a large set of data items. Frequent item set mining leads to the discovery of associations and correlation among items in large transactional or relational data sets.

Frequent items are finding more number of data mining algorithms used, Apriori algorithm of association rule which enumerate all of the frequent item sets. In this algorithm to make multiple passes over the database it is iterative approach. Dynamic frequent item set counting algorithm is enter the solve problem in Apriori. In this algorithm original database partition, single scan of database. So large update the repeated frequent items and association rule applying for candidate base items is finding. The following problems are identified in these techniques.

- Association mining rule applied for mostly candidate based.

- Apriori algorithm is finding the number of frequent items, search more database scan and pass take.
- DIC algorithms problem contains solve the Apiori algorithm base; it is large update the original dataset count. In this time large of database not frequently update the DIC algorithm.
- Original database is scan the repeated frequent items more compare the dataset. So waste timing and increasing the cost of finding market basket dataset.
- In parallelism two key performance issues are load balancing and synchronization. And another problem is that what happens if the size of the dataset is increased.
- No reduced the non frequent items. Because all items is maintains the last finding frequent item set. So main issue in store memory timing increasing, slow the all further finds frequent item sets.

IV. PROBLEM STATEMENT

Mining frequent item set in transaction databases, time-series databases, and many other kinds of databases has been studied popularly in data mining research. Most of the previous studies adopt an Apriori-like candidate set generation-and-test approach. However, candidate set generation is still costly, especially when exist a large number of patterns and/or long patterns. Further, dynamic item set counting algorithm, an extension to Apriori algorithm used to reduce number of scans on the dataset. It was alternative to Apriori Item set Generation .In this, item sets are dynamically added and deleted as transactions are read .It relies on the fact that for an item set to be frequent, all of its subsets must also be frequent, so we only examine those item sets whose subsets are all frequent. Both Apriori and DIC are based on candidate generation.

Previous algorithm is costly to handle a number of candidate's sets. The following example is explains:

If 1000 frequent 1-items set, the previous algorithms will need to generation more than 1000X 1000 length candidate and accumulated and test their occurrence frequent item set. This is the inherent cost of candidate generation, no matter what implementation techniques is applied.

It is tedious to repeatedly scan the database and check large set of candidates by item set matching, which is especially true for mining long patterns. Previous algorithm scans the database too many times, when the database storing a large number of data services, the limit memory capacity, the system I/O load, considerable scanning the database will be a very long time, so efficiency is very low.

In order to overcome the drawback inherited in previous algorithms, an efficient DIC- SCHEMA based mining method. In this method first classify the transaction items separate store classify schema and assume frequent items store the fuzzy schema.

Next step compare and finding frequent items between classify and fuzzy schema in this method used reduced more scanning the database and speed update of original database. A next non frequent item maintains dynamic schema, dynamic schema future partition of dynmaic1, dynamic2 and so on.

Next step compare and finding new frequent items, in this method used maintains less memory capacity. Final sequence update increase support percentage for original database.

V. ASSOCIATION RULE MINING USING CONFABULATION THEORY FOR FREQUENT ITEM SET MINING

CARM (Confabulation Inspired Association Rule Mining) approach uses a cogency inspired measure for generating rules. Cogency inspiration can lead us to more intuitive rules. Moreover, cogency-related computations only need pairwise item co occurrences, hence, rules can be found only by one file scan.

Rule mining is performed in two main phases: 1) knowledge acquisition and structure construction and 2) rule generation by confabulation and cogency measure. In this algorithm, only one item consequent association rules are generated, which means that the consequents of these rules only contain one item.

In CARM, items are considered as symbols. There are two modules in this system. Each module contains all items. Let I = {i1, i2, i3,..., im} be a set of items and T = {t1, t2, t3,..., tn} be a set of all transactions. L is a m×m matrix that stores knowledge link strength. First, knowledge links with weak strength are established between symbols of two modules. Then, it passes over database.

For each transaction ti ∈ T , all 2-subsets of ti are found. These subsets represent all co occurrences of items belonging to ti. Each 2-subset leads to strengthening the knowledge link between its items. In the simplest form, each element Lij of the matrix L represents the number of times that the link between item I and item j is strengthened. Therefore, principal diagonal stores the total number of times each item appears in database. During the database scan, for each transaction, all 2-subset are made and then their corresponding value in matrix L is incremented by one.

In ARM Phase, after finding all frequent 2-itemsets, the algorithm generates all rules using their support and confidence. In all computations, links with strength lower than a predefined minimum are discarded as uninteresting. All association rules in this algorithm are constructed using only the matrix L, so there is no need to mine frequent item sets list, which leads to speeding up the ARM process. The drawbacks are,

- Fixed transactions are considered.
- L matrix is not updated dynamically.
- MinCog parameter used which decides the association rule list to be added or not is fixed manually. It would be desirable to set it automatically using data set statistics.
- Forming new association rules by paying more attention to new data is not considered.

The proposed system includes all the existing system approach which covers CARM process. In addition, incremental information extraction is applied from the data sets. The new items from new transactions are found out and L Matrix size is incremented. Old L Matrix values are updated based on

old items found in new transactions. MinCog threshold is set based on the average link strength between items. The advantages can be listed as below:

- All the upcoming transactions are considered and applied in L matrix updating.
- MinCog parameter used which decides the association rule list to be added or not is set automatically using data set statistics.
- Forming new association rules by paying more attention to new data is considered.
- To reduce multiple scan of original database. Because previously select and set frequent items in the original database, so finding frequent items no repeated and newly enters the frequent items only scan and update datasets.
- To finding the frequent items based on fully transaction items, so reduced the more number of candidate generation.
- Effective algorithm and look for a balance between disclosure cost, computation cost and communication cost.
- Efficient and scalable methods for association rule mining should be developed in this algorithm.

VI. CONCLUSION

In this research work, we present the use of an association rule mining driven application is to manage market basket dataset that provide items with report regarding prediction of product purchase or sales trends and customer behavior. Our goal of the research is to find a new schema based rare frequent items for finding the rule of the market basket transactional dataset, which outperforms in terms of running time, number of database scan, memory consumption and the interestingness of the rules over the classical F-CARM algorithms. Our future work focuses on identifying frequent item sets with strong fuzzy items created techniques, update the original database increasing times, access the speed of processors implementation.

VII. REFERENCES

[1]. B. Nath, D. Bhattacharyya, and A. Ghosh, "Discovering association rules from incremental datasets," IJCSC, vol. 1, no. 2, pp. 433–441, 2010

[2] Y. Cao, H. He, and H. Man, "SOMKE: Kernel density estimation over data streams by sequences of self-organizing maps," IEEE Trans. Neural Netw. Learn. Syst., vol. 23, no. 8, pp. 1254–1268, Aug. 2012.

[3] P. Domingos and G. Hulten, "A general framework for mining massive data stream," J. Comput. Graphical Statist., vol. 12, no. 4, pp. 945–949, 2003

[4] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for on-demand classification of evolving data streams," IEEE Trans. Knowl. Data Eng., vol. 18, no. 5, pp. 577–589, May 2006

[5] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom, "Models and issues in data stream systems," in Proc. 21st ACM Symp. Principles Database Syst., 2002, pp. 1–16

[6] B. W. Silverman, Density Estimation for Statistics and Data Analysis. London, U.K.: Chapman & Hall, 1986

[7] J. DiNardo and J. L. Tobias, "Nonparametric density and regression estimation," J. Economic Perspectives, vol. 15, no. 4, pp. 11–28, 2001

[8] Y. Cao, H. He, H. Man, and X. Shen, "Integration of self-organizing map (som) and kernel density estimation (kde) for network intrusion detection," Proc. SPIE, vol. 7480, pp. 74800N-1–74800N-12, Sep. 2009

[9] P. N. Nohuddin, F. Coenen, R. Christley, C. Setzkorn, Y. Patel, and S. Williams, "Finding 'interesting' trends in social networks using frequent pattern mining and self organizing maps," Knowl. Based Syst., vol. 29, pp. 104–113, May 2012

[10] L. Gu, J. Li, H. He, G. Williams, S. Hawkins, and C. Kelman, "Association rule discovery with unbalanced class distributions," in Proc. 16th Austral. Joint Conf. Artif. Intell.,2003, pp. 221–232

[11] M. J. Heravi, "A study on interestingness measures for associative classifiers," M.S. thesis, Dept. Comput. Sci., Alberta Univ., Edmonton, AB, Canada, 2009

[12] Y. S. Koh and R. Pears, "Rare association rule mining via transaction clustering," in Proc. 7th Austral. Data Mining Conf., 2008, pp. 87–94

[13] Azadeh Soltani and M.-R. Akbarzadeh-T., Senior Member, IEEE, "Confabulation-Inspired Association Rule Mining for Rare and Frequent Itemsets", Ieee Transactions On Neural Networks And Learning Systems, Vol. 25, No. 11, November 2014 2053

AUTHORS BIOGRAPHY

Ms.T.Malathi Pursuing ME (CSE) degree in Shree Venkateshwara Hi-Tech Engineering College, Erode, India in 2014-2016 and BE(CSE) degree from Avinashilingam University for Women, Coimbatore, India in 2007-2011 .She is a Member of IEEE and Computer Society of India (CSI). She published 2 National Conferences, 1 Workshop. Her research interests include Software engineering, Cloud Computing and Data Mining.

Dr.T.Senthil Prakash received the Ph.D. degree from the PRIST University, Thanjavur, India in 2013 and M.E(CSE) degree from Vinayaka Mission's University, Salem , India in 2007 and M.Phil.,MCA.,B.Sc(CS) degrees from Bharathiyar University, Coimbatore India, in 2000,2003 and 2006 respectively, all in Computer Science and Engineering. He is a Member in ISTE New Delhi, India, IAENG, Hong Kong..IACSIT, Singapore SDIWC, USA. He has the experience in Teaching of 10+Years and in Industry 2 Years. Now He is currently working as a Professor and Head of the Department of Computer Science and Engineering in Shree Venkateshwara Hi-Tech Engineering College, Gobi, Tamil Nadu, and India. His research interests include Data Mining, Data Bases, Artificial Intelligence, Software Engineering etc.,He has published several papers in 17 International Journals, 43 International and National Conferences.

Mr.Arun.K Pursuing ME (CSE) degree in Shree Venkateshwara Hi- Tech Engineering College, Erode, India in 2014-2016 and BE(CSE) degree from the Shree Venkateshwara Hi- Tech Engineering College, Erode, India in 2008-2012.