International Journal of Computational Intelligence and Informatics, Vol. 5: No. 2, September 2015

An Approach to generate Cluster based Napped Associate Template Mining using Association Rules

Dr.R.U.Anitha
Muthayammal Engineering College, Rasipuram Namakkal,Tamilnadu
ruanithamca@gmail.com

Mrs.M.Menakapriya
Muthayammal Engineering College, Rasipuram Namakkal,Tamilnadu
menakaroshan@gmail.com

Mr.S.Nithyananth
Muthayammal Engineering College, Rasipuram Namakkal,Tamilnadu
nithyananthmca@yahoo.com

Abstract- **In this research we process a new technique called Napped Associate Template Mining (NATM) for concern forced Data Mining. It was used to find all the rules that capture the minimum support and minimum confidence constraints. In this proposed work, new template match technique to cluster association rules, based on the similar attributes, template matching clustering algorithm is used to cluster the rules. This work is used to combine more number of rules with a contingent value. Based on the contingent value, the result will be declared whether the rules or cluster or not.**

Keywords- Napped template, Association rules, clustering, knowledge group, cluster contraction.

I. INTRODUCTION

The historical data mining research contemplate more on the developing, demonstrating and pushing the use of the specialized algorithms and design [1].Data mining is the process of extracting desirable knowledge or interesting patterns from existing databases for specific purposes. Many types of knowledge and technology have been proposed for data mining. Among them, finding association rules from transaction data is most commonly seen. For instance, it is very interesting to Mining fuzzy rules is one of the best ways to summarize large databases while keeping information as clear and understandable as possible for the end-user.

Several approaches have been proposed to mine such fuzzy rules, in particular to mine fuzzy association rules [3]. However, we argue that it is important to mine rules that convey information about the order. Convey the idea of time running in rules, which is done in fuzzy sequential patterns. Regarding the evaluation results, the knowledge can be presented if the result is satisfactory, otherwise we have to run some or all of those processes again until we get the satisfactory result.

Combined mining is a technique for analyzing object relations and pattern relations, and for extracting and constructing actionable complex knowledge (pat-terns or exceptions) in complex situations [2]. Although combined patterns can be built within a single method, such as combined sequential patterns by aggregating relevant frequent sequences, this knowledge is composed of multiple constituent components (the left hand side) from multiple data sources which are represented by different feature spaces, or identified by diverse modeling methods. In some cases, this knowledge is also associated with certain impact (influence, action or conclusion, on the right hand side).

In this paper a set of categorical data for the clustering is collected. The frequent item sets are collecting using apriori algorithm. The item set are collected with the help of Weka tools. Using the item sets, the associated record sets from the datasets are collected. The collected record sets are clustered by grouping similar records. The attribute values are numeric or categorical data. This paper only considers the categorical data for this analysis.

II. RELATED WORKS

In this section, the association rule mining problem is presented in detail. Several issues in association rule mining have been elaborated together with classic algorithms. ARM aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data sources. The research activities in this area revolve around incorporating the mining capability into existing database technology, developing competent and scalable algorithms, handling user specific or domain specific constrains and Post processing of extracted patterns.

The AIS algorithm [4] was the first algorithm proposed to generate association rule. It focuses on improving the quality of databases together with necessary functionality to process decision support queries. The main disadvantage of the AIS algorithm is too many candidate itemsets that finally turned out to be small are generated, which needs more space and times. At the same time, this algorithm requires too many scans over the whole database to generate large itemsets.

The exploitation of monotonic property of the support of itemsets and confidence of association rules had been created a situation to enhance AIS algorithm and it was renamed as Apriori [5]. Apriori is a best improvement in the history of association rule mining. Apriori works efficiently during the candidate generation process for two reasons, Apriori employs a different candidates generation method and a new pruning technique. In the process of finding frequent item sets, Apriori avoids the effort of wastage of counting the candidate itemsets that are known to be infrequent. The candidates are produced by joining among the frequent itemsets level -wisely and are filtered according the Apriori property.

SETM (SET-oriented Mining of association rules) [6] was constantly outperformed by AIS. AprioriTid performed equivalently well as Apriori for lesser problem sizes. However, performance reduced twice slow when applied to big problems

The DHP (Direct Hashing and Pruning) algorithm is an effective hash-based algorithm for the candidate set generation [7]. Note that the DHP algorithm has two major features: One is its efficiency in generation of large item set and other is effectiveness in lessening on transaction data base size. Hence, a hash technique is very efficient in generating the candidate item sets, in particular for the large two item sets, thus really improving the performance bottleneck of the entire process.

An effective Direct Hashing and Pruning (DHP) algorithm [8] is proposed for mining the association rules. This algorithm employs effective pruning techniques to progressively reduce the transaction database size. DHP uses a hashing technique to screen the ineffective candidate frequent 2 item sets. DHP also avoids database scans in some passes as to reduce the disk I/O cost involved.

En et al. [9] introduced another novel hash-based approach for mining frequent itemsets over data streams. The algorithm compresses the information of all itemsets into a structure with a stable hash-based technique. This approach expertly summarizes the information of the whole data stream by using a hash table to estimate the support counts of the non-frequent itemsets and keeps only the frequent itemsets for speeding up the mining process.

This algorithm suggested by Brin et al. in 1997 [10] aimed at reducing the number of database scan by dividing the database into intervals of specific sizes. In effect, the algorithm reduces the number of database scans to 1.5 as contrasting to 3 scans required by level-wise priori approach. The algorithm is based on the overall principle of counting for item sets whenever it is optimal rather than having to wait for completion of the previous pass.

III. METHODOLOGY

We will focal point on telling the experiments planned to estimate the performance of the projected Data Structure Mining algorithm. At this time, Association ruling acting an important role. The purchasing of individual product when an additional product is purchased represents an association rule. The algorithm developed to present the distributed data at a very quick rate to the users engage flow of processing of data the same as follows

- Customer demands the data from the crossing point given. Data demanded is transferred to the proxy server, somewhere it is initially checked in the local database for simplicity of access, if the data is accessible, then provide to the user and occurrence of data is incremented, if not data is transfer to the Various Distributed databases using multithreaded atmosphere used for parallel processing.
- The variety of servers throws the number one winding up to the proxy server, where it is combined collectively to find the infrequent item set for the searched charge Item customer/Proxy Server mediator is acceptable to store the outcome close by so that Future search of the same value will not take longer instant.
- Proxy server mediator has been provide with the capability of setting Support threshold percent previous to handing out and also present the facility of searching for more than item at a time and in a quick rate of searching for particular value and more than one value a reduced amount of amount of time is preferred.

As a partitioning method, the k – means algorithm [Mac67] takes the input parameter, k , and partitions a set of n objects into k clusters with high intra-cluster similarity and low inter- cluster similarity. It starts by randomly selecting k objects as the initial cluster centroids. Each object is assigned to its nearest cluster based on the distance between the object and the cluster centroid. It then computes the new centroid (or mean) for each cluster. This process is repeated until the sum of squared-error (SSE) for all objects converges. The SSE is computed by summing up all the squared distances, one between each object and its nearest cluster centroid.

The data parallelism is used to divide the workload evenly among all processes. Data objects are statically partitioned into blocks of equal sizes, one for each process. Since the main computation is to compute and compare the distances between each object and the cluster centroids, each process can compute on its own partition of data objects independently if the k cluster centroids are maintained on all processes.

The algorithm is summarized in the following steps.

- Partition the data objects evenly among all processes;
- Select k objects as the initial cluster centroids;
- Each process assigns each object in its local partition to the nearest cluster, computes the SSE for all local objects, and sums up local objects belonging to each cluster;
- All processes exchange and sum up the local SSE's to get the global SSE for all objects and compute the new cluster centroids;
- Repeat (3) - (5) until no change in the global SSE

IV. RESULT AND DISSCUSSIONS

In this study, we take the data repository of bifocals. This data repository is based on the patient's eye problem[11][12]. It is an Attribute Relation File Format. Based on the data repository, we have various fields' attributes like age, bifocal prescription, and astigmation and tear production rate. The age attribute is classified into young, presbyopic, and pre-presbyopic. The bifocal prescription attribute have myope and hypermetrope. The astigmation have No and Yes. The tear production rates attribute has reduced and normal values. The class contact lenses have soft, hard and none. For fuzzy based combined pattern mining, first of all we find the fuzzy association rule. Then the fuzzy association rule merge with combined pattern mining

TABLE I. RECORD SETS OF CONTACT LENSES

S. No	AGE	BIFOCAL	ASTIGMATION	TEAR PRODUCTION RATE	CONTACT LENSES	CHOLESTROL
1	Young	Myope	No	Reduced	None	176
2	Young	Myope	No	Normal	Soft	155
3	Young	Myope	Yes	Reduced	None	173
4	Young	Myope	Yes	Normal	Hard	211
5	Young	Hypermetrope	No	Reduced	None	211
6	Young	Hypermetrope	No	Normal	Soft	165
7	Young	Hypermetrope	Yes	Reduced	None	176
8	Young	Hypermetrope	Yes	Normal	Hard	232
9	Pre-presbyopic	Myope	No	Reduced	None	162
10	Pre-presbyopic	Myope	No	Normal	Soft	174
11	Pre-presbyopic	Myope	Yes	Reduced	None	207
12	Pre-presbyopic	Myope	Yes	Normal	Hard	206
13	Pre-presbyopic	Hypermetrope	No	Reduced	None	181
14	Pre-presbyopic	Hypermetrope	No	Normal	Soft	159
15	Pre-presbyopic	Hypermetrope	Yes	Reduced	None	211
16	Pre-presbyopic	Hypermetrope	Yes	Normal	None	211
17	Presbyopic	Myope	No	Reduced	None	211
18	Presbyopic	Myope	No	Normal	None	203
19	Presbyopic	Myope	Yes	Reduced	None	211
20	Presbyopic	Myope	Yes	Normal	Hard	206

This data repository is applied in the Weka tool for finding frequent item sets with the help of association algorithm[13]. One of the popular algorithms for association rule called Apriori algorithm is applied in these data

repository by giving appropriate range of values. The item sets for the ten association rules are gathered and these frequent item sets are shown in the following figure with minimum confidence of 0.9.

TABLE II. ITEMSETS SATISFYING FOR ASSOCIATION RULE

Attribute 1	Attribute 2	Attribute 3	Attribute 4	Class
	Myope		Reduced	None
	Hypermetrope		Reduced	None
		No	Reduced	None
		Yes	Reduced	None
		No		Soft
			Normal	Soft
		No	Normal	Soft

This item sets are collected from association rules, that rules are matched with sample data sets. The item sets matched with each rule are stored in the separate table. Now the analysis will combine two or more rules. That is combining the data repository of the two association rules by a confidence value. Now collecting the data repository of the association rule 1 is matching the sample data repository. The resultant data repository is stored in a new table. Likewise for each association rule we are getting match record sets and stored in a separate table. The following table shows the record sets of the matching record of the association rule1 and rule 2 respectively[14].

TABLE III. DATA SET FOR ASSOCIATION RULE 1

AGE	SPECTACLE PRESCRIPION	ASTIGMATION	TEAR PRODUCTION RATE	CONTACT LENSES	CHOLESTEROL
Young	Myope	No	Reduced	None	176
Young	Myope	Yes	Reduced	None	181
Young	Hypermetrope	No	Reduced	None	174
Young	Hypermetrope	Yes	Reduced	None	173
Pre- presbyopic	Myope	No	Reduced	None	181
Pre- presbyopic	Myope	Yes	Reduced	None	183
Pre- presbyopic	Hypermetrope	No	Reduced	None	185
presbyopic	Myope	No	Reduced	None	189
presbyopic	Myope	No	Normal	None	178
presbyopic	Myope	Yes	Reduced	None	190
presbyopic	Hypermetrope	No	Reduced	None	176
presbyopic	Hypermetrope	Yes	Reduced	None	188

TABLE IV. DATA SET FOR ASSOCIATION RULE 2

AGE	SPECTACLE PRESCRIPION	ASTIGMATION	TEAR PRODUCTION RATE	CONTACT LENSES	CHOLESTEROL
Young	Myope	No	Reduced	None	176
Young	Myope	Yes	Reduced	None	181
Young	Hypermetrope	No	Reduced	None	174
Young	Hypermetrope	Yes	Reduced	None	173
Pre- presbyopic	Myope	No	Reduced	None	181
Pre- presbyopic	Myope	Yes	Reduced	None	183
Pre- presbyopic	Hypermetrope	No	Reduced	None	185
presbyopic	Myope	No	Reduced	None	189
presbyopic	Myope	No	Reduced	None	178
presbyopic	Myope	Yes	Reduced	None	190
Presbyopic	Hypermetrope	No	Reduced	None	176

The similar record sets are taken from two data sets and the counting is stored. The value is 11. The number of records in the first data set is 12. The number of records in the second data set is 11. By the equation 1,

Confidence value = 11/ (12+11-11)*100 = 91.66%.

V. CONCLUSIONS

A key reason for clustering rules is to obtain more concise and abstract descriptions of the data. In this research, the researcher considers only the non-numeric data, the main aim of this work is to reduce the number of groups or clusters. So the reduction of the drawback is rectifying using the iterative process. When number of iteration is increased, then less number of clusters is get. The most challenging problem in the data mining research and development is the mining complex data for complex knowledge. The fixed confidence or support value plays a main role in the cluster reduction analysis. The cluster reduction only depends on the fixed confidence of support value. As the fixed confidence or support value differs the number of cluster also differs. It means that whenever fixed confidence or support value reduces, it automatically reduces the number of clusters. This paper has presented the most comprehensive and a general approach called the combined mining using fuzzy concept, for discovering valuable knowledge in complex data. In future, the main aim of this research will find an accurate solution for checking low or previous order clustering to proceed for the higher order clustering. This research is extending to handle the numeric data.

REFERENCES

[1] Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., and Verkamo, A. I. "Finding interesting rules from large sets of discovered association rules". In Third International Conference on Information and Knowledge Management (CIKM'94), N. R. Adam, B. K. Bhar-gava, and Y. Yesha, Eds. ACM Press,pp 401-407, 1994.

[2] Koperski, K. and Han, J. "Discovery of spatial association rules in geographic information databases". In Proc. 4th Int. Symp. Advances in Spatial Databases, SSD,. Vol. 951. Springer-Verlag,pp 47-66,1995.

[3] F. Klawonn, R. Kruse, and H. Timm. "Fuzzy shell cluster analysis. In Learning, networks and statistics", pp 105–120. Springer, 1997.

[4] R.Agrawal, T.Imielinski, and A.Swami, 1993 "Mining Association Rules Between Sets Of Items In Large Databases", In proceedings of the ACM SIGMOD International Conference in Management of data, pp. 207-216

[5] Rakesh Agrawal and Ramakrishnan Srikant, "Fast Algorithms For Mining Association Rules In Large Databases", In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, Santiago, Chile, pp 487-499, Sep'1994.

[6] M. Houtsma, and Arun Swami, "Set-Oriented Mining for Association Rules in Relational Databases", IEEE International Conference on Data Engineering, pp. 25–33,1995.

[7] Park, J. S, Chen, M.S and Yu P. S, "An Effective Hash Based Algorithm For Mining Association Rules" , In Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, M. J. Carey and D. A. Schneider, Eds. San Jose, California, pp.175-186, 1995.

[8] Soo J, Chen, M.S, and Yu P.S, "Using a Hash-Based Method with Transaction Trimming and Database Scan Reduction for Mining Association Rules" , IEEE Transactions On Knowledge and Data Engineering, Vol.No.5. pp. 813-825,1997.

[9] EnTzu Wang and Arbee L.P. ChenData, " A Novel Hash-Based Approach For Mining requent Item-Sets Over Data Streams Requiring Less Memory Space", Data Mining and Knowledge Discovery, Volume 19, Number 1, pp 132-172

[10] S. Brin, R. Motwani, J.D. Ullman, and S. Tsur, "Dynamic Itemset Counting And Implication Rules For Market Basket Data", In Proceedings of the ACM SIGMOD, International Conference on Management of Data, volume 26(2) of SIGMOD Record, pp. 255–264. ACM Press, 1997.

[11] http://weka. wikispaces. com/Use+WEKA+in+your+Java+code#Classification-Building a Classifier.

[12] H.olmes, A. Donkin, I. H. Witten, WEKA: A Machine Learning Workbench, In Proceedings of the Second Australian and New Zealand Conference on Intelligent Information Systems, 357-361, 1994.

[13] Qiankun Zhao and Sourav S. Bhowmick," Association Rule Mining: A Survey",Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003116 , 2003.

[14] R. Agrawal, T. Imielinski, and A.N. Swami, "Mining Association Rules between Sets of Items in Large Databases," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 207-216, May 1993.