Synthesizing Model for Clustering Frequent Data Items in Multi-Database

R. Suganthi, P. Kamalakannan

Abstract - Mainly, most of the large organizations have numerous databases and they do process and transact over the multiple branch database. The important issue of the multi database is selecting the frequent items from various branch databases and forwarding the items to head quarters to take the decision among all kinds of patterns. Here global decision is important role in head quarter level and some steps are followed to take critical decision in top level. First step is synthesizing high frequency item set based on local item set. Second step is to measure the association* [13] *among various items listed under high frequency. And the accuracy level of data set is improved by using the synthesizing and clustering algorithm.

Keywords: Multi database, Synthesizing patterns, local pattern analysis, patterns

I. INTRODUCTION

In order to develop the business actions and industrial escalation, advanced and various new perceptions are needed to develop in information system and kinds of communication technologies. Exact decision will be taken through the spreading data over the multiple branch databases which are more critical issue to take decision in head quarters level. In these situations, traditional data mining techniques will not work properly due to the following reasons like high investment on software, hardware cost searching the individuality of local patterns and quantity of mixing data .This type of problem could be defeat by using multi data base mining [4] based on local pattern analysis. This technique will provide the frequent patterns among the databases and explained the technique of clustering local frequent item sets. Here the frequent patterns and association rules are the knowledge that we want to mine in such kind of scenario. Frequent patterns [2,9] are the patterns (Ex: item sets, subsequences (or) sub structure that appear frequently in a dataset. For ex., set of items, such as PC and memory card that appears frequently together in a transaction dataset is a frequent item set. Finding the association among the items [13] is the frequent patterns of the various applications.. This kind of association rule mining can be viewed as a 2 steps process. First, it search all frequent item sets. Second It Produces the Strong Association rules from the frequent item sets in all databases. Due to that traditional way of mining multiple data bases would not provide a fine solution to this kind of problem. In this situation, local pattern analysis could provide a solution. For getting the local patterns from the local data bases, traditional mining technique will be used. After wards, multi database mining would get patterns and forward into head quarters level. Then the patterns will be analyzed from different branch databases by using the synthesizing algorithm which makes the decisions to the related problem and the following are the exact definitions of various patterns. A *Local pattern* [1] **-** this style of pattern is based on only local database of each branch. And *High vote patters* are supported by mainstream of the branches or all branches of an interstate organization. These types of patterns replicate the widespread features among the branch database. *Exceptional patterns* [3] or *outlier patterns* have a higher support in few local branches. But zero support in other branches. From this perspective, organization would make some special policies for that branch alone. Using this model nonprofit item can get find easily and *suggested patterns* are supported by a few of the branches but these are lesser than the branches supporting the high vote patterns.

II. LITERATURE REVIEW

Wu & Zhang advocated [5] a model for synthesizing high frequency rules from multiple databases through weighting technique. It is a familiar method for aggregating information and needs to determine the weight of the data sources which is well thought-out as a first effort in synthesizing global patterns. Adhikari & Rao [6] have extended the local pattern analysis model and have introduced the notion of heavy association rules in their work. It is works on synthesized global supports which is greater than a user’s given threshold. Nedunchezian & Anbumani focused 2 issues namely data sources selection and selection of valid ruled for synthesizing. They have calculated the data source weight on the basis of 2 factors. One is that number of high frequency rules voted by data source and the data source size. Here threshold value is used to identify the candidate sources for synthesizing high frequency rules. And the following table summarizes the significant features of research work on the origin of synthesizing model strategy by using the clustering technique. Here the procedure of support equalization is used to reduce the total number of rules forwarded to the central head.

Manuscript Received on February 2015.

Asst Prof. Mrs. R. Suganthi , Department of Computer Applications, Valluvar college of Arts and Science, Karur, Tamil Nadu, India
Dr. P. Kamalakannan, Department of Computer Science, Government Arts College , Namakkal, India

Table 1. Analysis of research attempts in synthesizing model strategies

Author	Focusing of the Issues	Contribution
Animesh Adhikari	Synthesizing global exceptional patterns	Global Exceptional patterns which describe the Interesting individuality of few branches
Thirunavu -kkarasu Ramkumar	Multi level synthesis of frequent rules from different data sources	Multi level synthesis of local patterns on the basis of rule selection measure
Xindong Wu	Synthesizing high frequency rules from different data sources	Extracting high frequency rules from different data sources
Thirunavu -kkarasu Ramkumar	Correction factor of synthesizing global rules [10]	Model for synthesizing Correction factor of synthesizing global rules in multi data base mining
Animesh Adhikari	Clustering local frequency items in multiple databases	Designed algorithm for synthesizing supports of such item sets using clustering technique

III. IMPLEMENTATION DETAILS

We have evaluated the effectiveness of our synthesizing approach by conducting various experiments. The results of 10 databases of our studies are discussed further. The idea of the project basically resolves around the concept of synthesizing [12] and clustering the local frequency items in multiple data bases. it has 2 major steps.

1. Synthesize high frequency item sets based on local item sets.
2. Clustering the frequent items based on associations.

IV. SYNTHESIZING HIGH FREQUENT ITEMS

In order to improve the quality of global patterns from multiple databases, pipelined feedback technique has been implemented to mine the multiple large databases and synthesizing algorithm which is used to find association among items in a data bases and it returns highly accurate patterns .

Algorithm:

1. Store all the local item sets into array
2. Sort the item sets based on item set attribute.
3. Add same size of all multi branch data bases into Transactions
4. Keep track of the number of synthesized item sets and the Current item sets.
5. Determine whether an item set has high frequency Increased by 1 or not
6. Calculate the synthesized association among items for Each synthesized item set of size greater than 1

V. LOCAL FREQUENCY ITEMS OBTAINED BY CLUSTERING TECHNIQUE

Measure of similarity [14] is the existing technique for clustering local frequency items. first it would compute the relevant databases [15] along with similarity items in a database. For finding the non trivial partition of local frequency items in multi databases, synthesizing high frequency item size has to be calculated in synthesizing algorithm that value should be greater than 1.

Algorithm:

1. Synthesized elements should be arranged on the basis of Associating the items in a data set
2. Synthesized high frequency item set forms a single ton Class.
3. Partition has to be formed based on synthesized association Among items
4. Accumulates all the items.
5. Check the mutual exclusiveness among the previous and Present frequent item set.
6. It finds the best partition and clustering the local frequency items from all the databases.

VI. EXPERIMENTAL RESULTS AND SCALABILITY

We have carried out several experiments on Net Beans IDE 8.0 software and study the effectiveness of our approach. All the experiments have been implemented on Intel Pentium 2.13 Dual core processor with 2 Gb RAM and 32 bit operating system. The experimental results are printed on synthetic and real data set. Each database item is indicated by an integer to perform experiments more conveniently. The synthetic dataset SD 1,00,000 [11]obtained from an Belgian retail supermarket store and specified the data set characteristics.

The details of dataset SD1, 00,000 is the following,

Number of transactions – 1,00, 000
Total data set - 10
Average length of transactions - 16.4
Divided datasets - D1,D2,… .D10
Average frequent item sets -153.10
Number of items in each data set- 10,000
The input databases obtained from SD 1,00,00(S) is Given as follows.,

D1 ={1,2,3,4,5,6,7,8,9,10…..20}
D2 ={ 8,12,14,15,16,17,18}
D3 ={1,5,8,,10,12,14,15,16,17,18}
D4 ={ 1,5,8,10,12,15,16,17,18}
D5 ={1,7,8,10,12,13,14,15,16,17,18}
D6

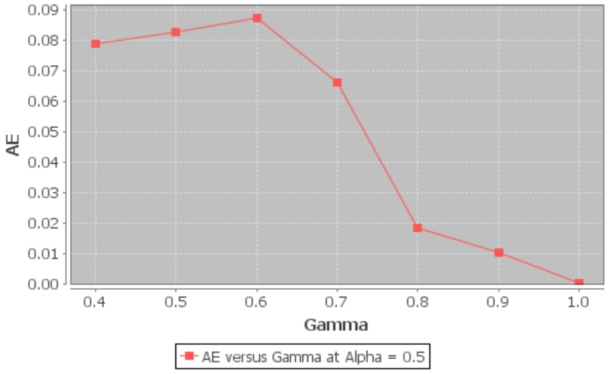
={1,5,8,9,10,12,15,16,17,18}
D7 ={1,8,9,10,12,16,17,18,19}
D8 ={1,8,10,12,16,17,18,19 }
D9 ={ 8,10,12,14,15,16,17,18}
D10 ={ 2,7,10,12,13,16,17,18}

After the experimental work, maximum error and average error has to be measured using the following formula.

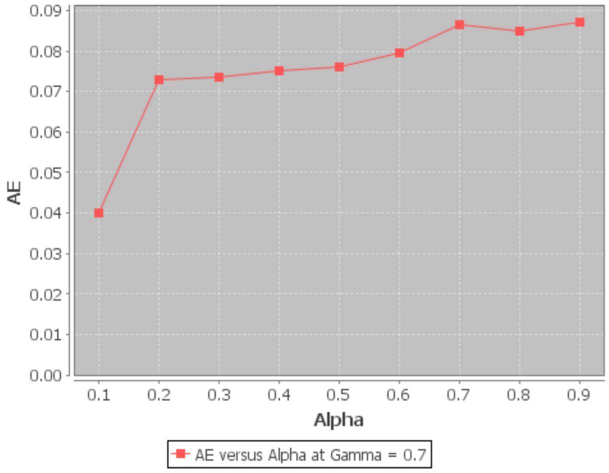
1. $AE\ (D,\alpha,\gamma) = 1/M \sum_{i=1}^{M} |\ SS\ (Xi, D) - S(Xi\ ,\ D)$

2. $ME\ (D,\alpha,\gamma) = \text{Maximum}\ \{|\ SS\ (Xi, D) - S(Xi\ ,\ D)$
$I = 1,2,\text{--}\ m\ \}.$

The following graph mentioned the synthesized time and clustering time based on the alpha and gamma values provided
Alpha = 0.5 (Fixed)
Gamma = 0.4 to 1



Here α is user defined level of min support. more experiments are conducted to study the behavior of average error over different α S based on the clustering time of an experiment will be increased because of increasing the number of databases . And the other graph value is,
Gamma = 0.7 (fixed)
Alpha = 0.1 to 0.9



After executing the proposed algorithm, the performance is good with the result of SD(1,00,000) data sets and the partitioning of the 10 data sets along with α = 0.1 and γ = 0.7 is produced. And the performance is high with the value of 0.399.But in the existing method the performance was 0.251

VII. CONCLUSION

The clustering of local frequent items is an important component of multi database mining system. It reduces the cost of searching relevant information for many problems. We present an efficient solution to this problem in 2 possible ways. First step is to synthesize the association among items. Second step is to finding the best non trivial partition of local frequency items in multiple data bases. It has been observed that an existing clustering technique might cluster local frequency items at a minimum level even the high level associated items in all data bases. Thus the proposed algorithm is clustering the local frequent items at highest level.

REFERENCES

1. T. Ramkumar, S. Hariharan, S. Selvamuthukumaran, A survey on mining multiple data sources, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 3 (1) (2013) 1–11.
2. Han/Kamber/Pei, Tan/Steinbach/Kumar, and Andrew Moore , Data Mining Techniques Classification and Prediction 1997 : 33-70
3. A. Adhikari, Synthesizing global exceptional patterns in different data sources, Journal of Intelligent Systems 21 (3) (2012) 293–323
4. A. Adhikari, P. Ramachandrarao, W. Pedrycz, Developing Multi-Database Mining Applications, Springer-Verlag, London, 2010.
5. Wu X, Zhang S. Synthesizing high-frequency rules from different data sources. *IEEE Trans Knowl Data Eng* 2003, 15:353–367..
6. Adhikari A, Rao PR. Synthesizing heavy association rules from different real data sources. *Pattern Recognit Lett* 2008, 29:59–71
7. Nedunchezhian R, Anbumani K. Post mining– discovering valid rules from different sized data sources. *Int J Inf Technol* 2006, 3:47–53.
8. A. Adhikari, Clustering local frequency items in multiple databasess, Journal of Information Science 237 (2013) 221–241
9. J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan
10. Kauffmann Publishers, 2001.
11. Ramkumar T, Srinivasan R. The effect of correction factor in synthesizing global rules in a multi-database mining scenario. *J Appl Comput Sci* 2009, 3:33– 38
12. Frequent Itemset Mining Dataset Repository. http://fimi.cs.helsinki.fi/data/
13. Kum HC, Chang JH, Wang W. Sequential pattern mining in multidatabases via multiple alignment. *Data Min Knowl Discov* 2006, 12:151–180:
14. Savasere A, Omiecinski E, Navathe S. An efficient algorithm for mining association rules in large databases. In: *Proceedings of the Twenty First International Conferences on Very Large Data Bases*. Zurich, Switzer-land; 1995, 432–444
15. Lenca P, Meyer P, Vaillant B, Lallich S. On selecting interestingness measures for association rules: user ori-ented description and multiple criteria decision aid. *Eur J Oper Res* 2008, 184:610–626
16. Liu H, Lu H, Yao J. Toward multi-database min-ing: identifying relevant databases. *IEEE Trans Knowl Data Eng* 2001, 13:541–553