A Novel Progressive Sampling based Approach for Effective Mining of Association Rules

V.Umarani
Asst Professor, Dept of Computer Science
Sri Ramakrishna College of Arts and
Science for Women,
Coimbatore, India

Dr.M.Punithavalli
Director and Head, Dept of computer science,
Sri Ramakrishna College of Arts and
Science for Women,
Coimbatore, India.

ABSRACT

Mining Association Rules from huge databases is one of the important issue that need to be addressed. This paper presents a new sampling based association rule mining algorithm that uses a progressive sampling approach based on negative border and Frequent pattern growth (FP Growth) algorithm for finding the candidate item sets which ultimately shortens the execution time in generating the candidate itemsets. Experimental results reveals that the propsed approach is significantly more efficient than the Apriori based sampling approach.

Keywords: Apriori, Negative border, FP-Growth, Sampling, Temporal Characteristics

I. INTRODUCTION

Association Rule Mining finds relations among data items in transactions of a huge database. Association rules are frequently used by retail stores to assist in marketing, advertising, floor placement and inventory control. Although they have been used for other purposes as well, including predicting faults in telecommunication networks. Association rules are used to show the relationship between data items.

This research focuses on finding frequent patterns quickly by sampling of transaction database progressively. The method uses the combination of FP Growth and our earlier approach[14] to fasten association rule mining process in huge databases.

2. RELATED WORKS

A number of studies were conducted to propose efficient methods for mining association rules by reducing either the CPU computation time or the disk access overhead [4,11].Some studies considered the usage of sampling techniques for reducing the processing overhead [2,16]. Most of the prior works on sampling have concentrated on speeding up the phase by running a frequent itemset mining algorithm only on a small sample of the database [15]. Chiefly, researchers have evaluated the viability of using sampling [10] to reduce the dataset size. While such methods have shown quite a lot of promise it has been observed by several researchers [15,17,] that it is often very difficult to quantify, apriori, the quality of the results obtained for a given sample size [12], necessitating novel and more effective sampling-based association rule mining algorithms to foster better mining results.

Several researches are available in the literature for sampling-based association rule mining. A brief review of some of the significant researches is presented here.

Basel A. Mahafzah *et al.* [3] have presented a parameterized sampling algorithm for association rule mining. The algorithm extracts sample datasets based on three parameters: transaction frequency, transaction length and transaction frequency-length.To evaluate its performance and accuracy, a comparison against a two-phase sampling based algorithm [5] was performed using real and synthetic datasets. The experimental results showed that the proposed sampling algorithm in some cases outperformed two phase sampling algorithm in terms of accuracy. Cai-Yan Jia and Xie-Ping Gao [6] have presented an adaptive, on-line, fast sampling strategy which was inspired by MRA (Multi-Resolution Analysis) and Shannon sampling theorem, for quickly obtaining acceptably approximate association rules at appropriate sample size. Both theoretical analysis and empirical study have showed that the sampling strategy can achieve a very good speed-accuracy trade-off.

Venkatesan T *et al.* [15] have presented a comprehensive theoretical analysis of the sampling technique for the association rule mining problem. The sampling based technique was used to solve frequent itemset mining and association rule mining problems using a sample whose size is independent of both the number of items and the number of transactions. Thus, the possibility of speeding up the entire process of association rule mining for massive databases by working with a small sample while retaining any desired degree of accuracy was established. Zhao *et al.* [17] have proposed a hybrid theoretical bound of sample size for frequent itemsets discovering and association mining. By combining the additive error bound and the multiplicative error bound, the proposed bound makes the theoretical sample size to be much less than traditional Chernoff bounds. Theoretical analysis showed that the sample size is about an order of magnitude smaller than the traditional Chernoff bounds. The experiment results validated the effectiveness of the proposed bounds.

Chuang K *et al.* [7] have presented a progressive sampling algorithm, called Sampling Error Estimation (SEE), which aims to identify an appropriate sample size for mining association rules. Sampling error estimation has two advantages over previous works in the literature. First, it is highly efficient because an appropriate sample size can be determined without the need of executing association rules. Second, the identified sample size is very accurate, meaning that association rules can be highly efficiently executed on a sample of this size to obtain a sufficiently accurate result. This attributed to the merit of sampling error estimation for being able to significantly reduce the influence of randomness by examining several samples with the same size in one database scan. As validated by experiments on various real data and synthetic data, sampling error estimation can achieve very prominent improvement in efficiency.

Bin Chen Exelixis *et al.* [5] have introduced FAST (Finding Associations from Sampled Transactions), a two-phase sampling-based algorithm for discovering association rules in large databases. In Phase I, a large initial sample of transactions was collected and used to quickly and accurately estimate the support of each individual item in the database. In Phase II these estimated supports were used to either trim "outlier" transactions or select "representative" transactions from the initial sample, thereby forming a small final sample that more accurately reflects the statistical characteristics of the entire database. In an empirical study, FAST was able to achieve 90-95% accuracy using a final sample having a size of only 15-33% of that of a comparable random sample. The sampling technique can be used in conjunction with almost any standard association-rule algorithm, and can potentially render scalable other algorithms that mine "count" data.

3. WHY FP GROWTH TREE ALGORITHM?

Apriori[1] and its dialect algorithms find associations in a direct manner by successively growing multi-itemsets starting from 1-itemset that occurs frequently above some threshold value. However, they require multiple iterations to prune candidate itemsets which ultimately results in long computing time.

FP-growth (frequent pattern) tree[5] is an effiecient algorithm for finding frequent patterns in transaction database. FP-growth (frequent pattern growth) uses an extended prefix-tree structure to store the database in a compressed form. FP-growth adopts a divide and conquer approach to decompose both the mining tasks and the databases. It uses a pattern fragment growth method to avoid costly process of candidate generation and testing used by Apriori. As Mining is based on the tree structure, FP-tree is significantly more efficient than Apriori.

Steps of FP tree Algorithm are:

1) Construct conditional patternbase for each node in FP-tree.
2) Construct conditional FP-tree from each conditional pattern base.
3) Recursively mine conditional FP-trees and grow frequent patterns.

Hence in the proposed work, FP growth tree algorithm is used to overcome the main memory limitation which ultimately results in efficient mining of association rule from huge databases.

4. PROPOSED WORK

This section describes the progressive sampling based approach for effective association rule mining. The important aspect is to choose an initial sample of data from database. Most of the existing algorithms choose the samples in a random manner without taking into account any aspects of the database and it results in difficulty in obtaining samples.Hence for obtaining an optimal sample, our previous work[14] considers the following:

a) the temporal characteristics of the database

b) progressive sampling based on negative border.

In our Earlier approach[14] we have used traditional Apriori algorithm for generating candidate itemset. But in this proposed approach we have made use of FP tree algorithm for generating candidate itemsets which ultimately shortens the execution time while generating candidate itemsets.

Following are the steps involved in Proposed approach:

1. Selection of initial sample S_i using systematic sampling of size ‘n’ from the original database(D) by making use of temporal characteristics.
2. Mining of frequent itemsets using FP tree growth algorithm and generate the negative border.
3. Sort the negative border itemsets according to their support level.
4. Select the midpoint itemset from the sorted negative border.
5. Determine the support of midpoint itemset by single scan on the remaining database D1 of size ‘(d-n)’, where $D1=D-S_i$.
6. If the support of midpoint itemset is less than the user specified support, the chosen sample is known as optimal sample. Otherwise, the sample size ‘n’ is increased and steps 2-5 are performed repeatedly until an optimal sample is selected.

Input:

Transaction database D, Size of initial Sample in %, increment of sample size in %, minsupp, minconf.

Output:

Optimal sample size, Association rules.

5. RESULTS AND DISCUSSION

This section describes the results obtained on experimentation of the proposed sampling approach for ARM. The proposed method is implemented in (jdk1.6). We now describe the experiments in order to assess the practical feasibility of using samples for finding frequent itemsets. The proposed work is evaluated on the synthetic dataset which contains in it, 30k transactions and 11 itemsets.

Apriori, the most renowned algorithm has been chosen for evaluating the performance of our proposed approach. The notion of model accuracy for a particular dataset varies sensitive to relevant interaction parameters(e.g. support, confidence,important items to the user) as well as inherent properties of the dataset in question. Here we describe the accuracy as the number of association rules discovered from different support thresholds by the proposed approach in comparision to Apriori algorithm rather than the similarity against individual rules.

We have chosen the parameters such as i) Time complexity and ii) Accuracy for comparing our earlier approach with Apriori. In addition we have obtained Optimal sample size for mining of association rules.

A number of association rules mined by Apriori, our earlier approach and our proposed approach are illustrated in Fig 1. The result shows that there has not so many rules left mined by the proposed approach when compared to Apriori and obviously with increased thresholds (more significant patterns), the proposed approach almost achieves the model accuracy of the classical Apriori algorithm.

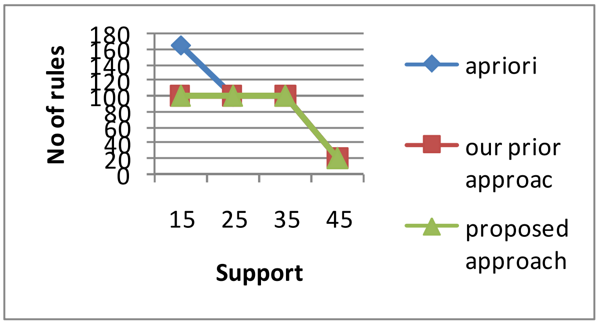

Fig.1 Accuracy graph of data set

An important measure that illustrates the effectiveness of the sampling-based approaches is the timing incurred to complete ARM.The computation time required for mining the Association rules from the synthetic dataset is plotted as a graph shown in Fig 2. Vividly, we can see an appreciable reduction in timing required for ARM using Apriori and proposed approach.

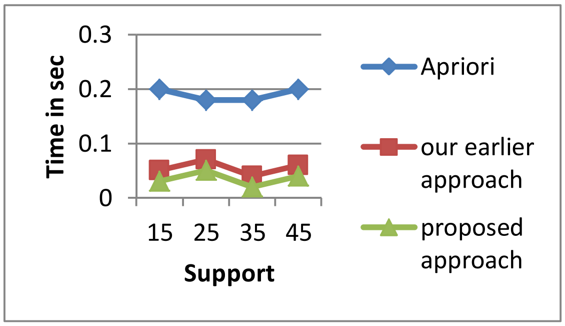

Fig.2 Time graph for dataset

Optimal Sample Size: One of the important scenarios in the sampling approach is to find the optimal sample size while mining association rules. This concludes that the association rules can be mined within this optimal sample rather than the entire database.

The Optimal sample size obtained for synthetic data is given in Table 1.

Dataset	
Support	**Optimal Sample Size**
15	**20**
25	**20**
35	**20**
45	**20**

Table 1. Optimal sample size of Synthetic dataset.

6. CONCLUSION

This paper projected the impact of using FP growth tree along with sampling technique. The performance study shows that utilizing FP growth tree algorithm along with sampling would yield better result in efficient and scalable mining of association rules from huge databases. Also the proposed approach is significantly more efficient than Apriori based Sampling sampling approach.

7.REFERENCES

[1] R.Agarwal and R.Srikant,"Fast algorithms for mining association rules". In Proc. VLDB Conf., pp 487-499.

[2] R.Agrawal, T.Imielinski, and A.Swami, "Mining association rules between sets items in large databases", in proceedings of the ACM SIGMOD Int'l Conf. on Management of data, pp. 207- 216, 1993.

[3] Basel A. Mahafzah, Amer F. Al-Badarneh and Mohammed Z. Zakaria "A new sampling technique for association rulemining," in Journal of Information Science, Vol. 35, pp. 358-376, 2009.

[4] S. Brin, R. Motwani, J. D. Ullman and S. Tsur, "Dynamic Itemset Counting and Implication Rules for Market Basket Data," Proc. ACM SIGMOD, 1997, pp. 255-264.

[5] B. Chen, P.Haas, and P.Scheuermann," A new two phase sampling based algorithm for discovering association rules",SIGKDD, 2002.

[6] Cai-Yan Jia and Xie-Ping Gao, "Multi- scaling sampling: an adaptive sampling method for discovering Science and Technology archive, Vol. 20, pp. 309-318, 2005.

[7] Chuang K, Chen M, Yang .W,"Progressive Sampling for Association Rules based on Sampling Error Estimation", Lecture notes in computer Science, Vol. 3518, pp. 505-515,2005.

[8] J.Han, J.Pei, and Y.Yin,"Mining frequent patterns without candidate generation", SIGMOD,2000.

[9] Hannu Toivonen, "Sampling Large Databases for Association Rules", Proceedings of the 22nd International Conference on Very Large Data Bases, pp: 134 - 145, 1996 SIGMOD,2000.

[10] Klaus Julisch," Data Mining for Intrusion Detection -A Critical Review" in proc. Of IBM Research on application of Data Mining in Computer security, Chapter 1 , 2002.

[11] J. S. Park, M. S. Chen, and P. S. Yu, “An Effective Hash based Algorithm for mining association rules,” Proc. ACM SIGMOD Conf Management of Data, May, 1995.

[12] Parthasarathy, S., "Efficient progressive sampling for association rules", IEEE International Conference on Data mining, pp: 354- 361, 2002.

[13] Raymond Chi-Wing Wong, Ada Wai- Chee Fu, "Association Rule Mining and its Application to MPIS", 2003.

[14] V.Umarani, M.Punithavalli,” On developing an effectual progressive sampling based approach for Association Rule Discovery”, In the proceedings of 2nd IEEE ICIME Int’l conference on Information and Data Management”, Chengdu,China.

[15] Venkatesan T. Chakaravarthy, Vinayaka Pandit and Yogish Sabharwal, "Analysis of sampling techniques for association rule mining," In Proceedings of the 12thInternational Conference on Database Theory, Vol. 361, pp. 276-283,2009.

[16] M. J. Zaki, S. Parthasarathy, W. Li, and M. Ogihara,“Evaluation of Sampling for Data Mining of Association Rules,” Technical Report 617, CS Dept., U. Rochester, May 1996.

[17] Y. Zhao, C. Zhang and S. Zhang, “Efficient frequent itemsets mining by sampling,” Proceedings of the fourth International Conference on Active Media Technology (AMT), pp. 112-117,