Sampling based Association Rules Mining- A Recent Overview

V.Umarani,

Lecturer,
Dept of Computer Science,
Sri Ramakrishna College of Arts and Science for Women,
Coimbatore-44,
India.

Dr.M.Punithavalli,

Director and Head of the Department,
Dept Of Computer Science,
Sri Ramakrishna College of Arts and Science for Women
Coimbatore-44
India.

Abstract

Association rule discovery from large databases is one of the tedious tasks in datamining.The process of frequent itemset mining, the first step in the mining of association rules, is a computational and IO intensive process necessitating repeated passes over the entire database. Sampling has been often suggested as an effective tool to reduce the size of the dataset operated at some cost to accuracy. Data mining literature presents with numerous sampling based approaches to speed up the process of Association Rule Mining(ARM).Sampling is one of the important and popular data reduction technique that is used to mine huge volume of data efficiently. Sampling can speed up the mining of association rules. In this paper, we provide an overview of existing sampling based association rule mining algorithms.

Keywords: Datamining, sampling, Association rule mining, data reduction technique, Frequent patterns

I. INTRODUCTION

The volume of electronically accessible data in warehouses and on the internet is growing faster then the speedup in processing times predicted by Moore's law[32].Scalability of mining algorithms is therefore a major concern. Classical mining algorithms that require one or more passes over the entire database can take hours or even days to execute and in the future this problem will still worsen. One approach to the scalability problem is to exploit the fact that approximate answers often suffice and execute mining algorithms over a 'synopsis' or 'sketch'. The computation of many synopses is been proposed in the literature [7] requires one or more expensive passes over all the data, so that the use of synopses may still fail to adequately address the scalability problem unless the cost of producing the synopses is amortized over many queries.

Using a sample of the data as the synopses is the popular technique that can scale well as the data grows. Besides having desirable scaling properties, sampling is also suited to interactive exploration of massive data sets.

Recent work in the area of approximate aggregation [4] processing shows that the benefits of sampling are most fully realized when the sampling technique is tailored to the specific problem at hand.

A. Association Rule Mining and its Applications

Association rule mining, one of the important technique which aims at extracting interesting correlations, frequent patterns, associations or casual structures among set of items in the transaction databases or other data mining repositories. Among the areas of data mining, the problem of deriving associations has received a great of attention. The problem was formulated by Agarwal et al [3] in 1993 and is often referred to as market-basket analysis. Association rule mining is to find out association a rule that satisfies the predefined minimum support and confidence from a given database. The problem is usually decomposed into two sub problems. One is to find those items sets whose occurrences exceed a predefined threshold in the database: those item sets that are called frequent or large item sets. The second problem is to generate association rules from those large item sets with constraints of minimal confidence. Suppose one of the large item sets is L_k, L_k = $\{I_1,I_2,\ldots I_k\}$, Association rules with these item sets are generated in the following way: The first rule is $\{I_1,I_2,\ldots..I_{k-1}\} \Rightarrow \{I_k\}$, by checking the confidence this rule can be determined as interesting or not. Then other rule are generated

by deleting the last items in the antecedent and inserting it into the consequent, further the confidence of the new rules are checked to determine the interestingness of them. Those processes iterated until the antecedent becomes empty. Since the second problem is quite straight forward, most of the researches focus of the first sub problem.

The first sub problem can be further divided in to two sub problems: candidate large item set generation process and frequent item sets generation process. We call those item sets whose support exceed the support threshold as large or frequent item sets, those item sets that are expected or have the hope to be large or frequent are called candidate item sets.

Association rule mining is one of the important issues in data mining and there is a high demand in industry as extraction of association rules is directly related to sales. Association rule is also applied in telecommunication networks. Association rules also finds application in various areas like telecommunication networks, market and risk management, inventory control

.

B. Sampling and Association Rule Mining

The importance of sampling for association rule mining has been recognized by several researchers [5, 21, 22, 29, 30]. The usual approach is to take a portion of database randomly of a previously determined size and then calculate the frequency of item sets over the sample using a lower minimum support threshold σ' that is slightly smaller than the user – specified minimum support σ. Also the computational cost of association rule mining can be reduced in four ways:

- By reducing the number of passes over the data base[1,15,37]
- By sampling the database
- By adding extra constraints on the structure of patterns [1519, 12, 25].
- Through parallelization.[20,21,22,23]

The discussion here is limited to sampling. The rest can be referred from [15].The task of mining association rules is usually performed in transactional or relational databases, to derive a set of strong association rules may require repeated scans through the database. Therefore, it can result in huge amount of processing when working on a very large database. Many efficient algorithms can significantly improve the performance in both efficiency and accuracy of association rules [15] .However; sampling can be a direct and easy approach to improve the efficiency when accuracy is not the soul concern.

II. RELATED WORKS

Toivonen et al [28] presented an Association rule mining algorithm using sampling. The approach can be divided into 2 phases. In phase I, a sample of the database is obtained and all associations in the sample is obtained and all associations in the sample of the database are found. These results are then validated against the entire database. To maximize the effectiveness of the overall approach the author makes use of lowered minimum support on the sample. Since the approach is dependent (probabilistic) on the sample containing all relevant associations, not all the rules may be found in the first pass. These associations that were deemed not frequent in the sample but were actually frequent in the entire dataset are used to construct the complete set of associations in phase 2.

Mohammed et al [20] reviews and proposed that random sampling of transactions in the database is an effective way for finding association rules. They have the following contributions i. Sampling can reduce i/o cost by drastically shrinking the number of transactions to be considered and ii. Sampling can provide greater accuracy with respect to the association rules. They have shown that sampling can speed up the mining process by more than a order of magnitude.

Parthasarathy [22] proposed an efficient method to progressively sample for association rules. his approach relies on a novel measure of model accuracy(self-similarity of associations across progressive samples),the identification of a representative class of frequent item sets that mimic(extremely accurate) the self-similarity values across the entire set of association and an efficient sampling methodology that hides the overhead of obtaining progressive samples by overlapping it with useful computation.

Chen et al [8] presented a novel two phase sampling algorithm for discovering association rules in large databases .These algorithm has 2 phases. In phase I, a large initial sample of transactions is collected and used to quickly and accurately estimate the support of each individual item in the database. In phase II, these estimated supports are 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(i.e. itemset supports) of the entire databases. The expensive operation of discovering association rules is then performed on the final sample.

Bronnimann et al [6] explored another sampling algorithm called epsilon approximation: sample enabled (EASE). Unlike FAST [8] which obtains final sub sample by quasigreedy descent, EASE uses Epsilon approximation methods to obtain the final sub sample by process of repeated halving. This algorithm can process transactions on the fly,i.e a transaction needs to be examined only once to determine whether it belongs to the final sub sample.

Surong et al presented another algorithm called EASIER [25] which is an extension to EASE [6] in two ways. 1) EASE is a halving algorithm i.e. to achieve the required sample ratio it starts from a suitable initial large sample and iteratively halves. EASIER on the other hand, does away with the repeated halving by directly obtaining the required sample ratio in one iteration. 2) EASE was shown to work on IBM Quest dataset which is a categorical count data where as EASIER in addition to count data it was also shown on Continuous data such as Color Structure Descriptor (CSD) of images.

Chuang et al [9] presented another progressive algorithm called Sampling Error Estimation (SEE) which aims to identify an appropriate sample size for mining association rules .SEE has two advantages 1. SEE is highly efficient because an appropriate sampling size can be determined without the need of executing association rules. 2. The identified sample size of SEE is very accurate (i.e.) the association rules can be highly efficiently executed on a sample of this size to obtain a sufficiently accurate result.

Wontae et al [34] presented a new algorithm called IFAST that uses two phase sampling [8] algorithm for shortening the execution time at the cost of precision of the mining result. Previous FAST [8] algorithm has the weakness in that it only considered the frequent 1-itemsets in trimming/growing phase. thus it did not have ways of considering multi-item sets including 2itemsets.IFAST algorithm reflects the multi-item sets in sampling transactions. It improves the mining results by adjusting the counts of both missing item sets and flase itemsets.

Venkatesan et al [29] proposed a different view of analyzing the quality of solution by theortical framework.their contributions is twofold. First ,the notions of e-close frequent item set mining and e-close association rule mining that help assess the quality of solutions obtained by sampling .secondly, the frequent itemset mining and association rule mining problem can be solved satisfactorily with a sample size that is independent of both the number of transactions size and number of items. It has also been established that it is possible to speed up the entire mining process of association rule mining for massive databases by working with a small sample size while retaining any desired degree of accuracy. Their work also gives a comprehensive explanation for well known empirical success of sampling for association rule mining.

Basel et al [5] recently presented a parameterized sampling algorithm for association rule mining. This algorithm extracts sample datasets based on three parameters: transaction frequency, transaction length and transaction frequency length and it empirically shown that it achieves 98% accuracy which outperform two-phase algorithm [6].

Our earlier work [37] we presented a progressive sampling-based approach for mining association rules from massive databases. This approach aims to fasten and produce acceptable accuracy in association rule mining. The concept of progressive sampling has been made use of in the proposed approach for identifying a fitting sample of large database. Here the Sample selection is based on temporal characteristics of the original database. Progressive Sampling is done on estimated negative border. The proposed approach is likely to yield considerable reduction in computational time with some cost to accuracy (optimality between accuracy and time).

III.CONCLUSION

Data reduction is concerned with reducing the volume of data while retaining its essential characteristics. As such, sampling provides a general approach which scales well and offers more flexibility than merely tracking count statistics. Moreover, the sample can better be used for training purpose or further statistical analysis. For the full benefit of sampling, however, it is best to tailor a sampling procedure to the problem at hand. Sampling is one of the important techniques to increase the efficiency of association rule mining [22]

For a specific data mining task under specific data set, one sampling strategy may work better than others in terms of accuracy or efficiency [15]. There fore it is necessary to study how different strategies are in a specific data mining task given specific data sets in order to provide users a set of guidelines for them to make decisions on which context it will be more suitable to use which sampling strategy.

References:

1. Agarwal, .R. Aggarwal, C. and Prasad V., A tree projection algorithm for generation of frequent itemsets.In J.Parallel and Distributed Computing.2000.
2. Agarwal. R. Srikant.R: Fast algorithms for mining association rules. In proc Int'l Conf on VLDB (1994).
3. Agarwal. R. Srikant.R: Fast algorithms for mining association rules. In proc Int'l Conf on VLDB (1994).
4. Acharaya, s.; Gibbons, P.B.; and Poosala, V.2000.Congressional samples for Approximate Answering of Group-By Queries. In Proceedings of the ACM SIGMOD International Conference on Management of Data.
5. Basel et al,"a new sampling technique for Association rule mining", Journal of information science ,June 2009,vol 35, pp 358-376
6. Bronnimann .H,Chen B, Dash M, Haas,.P,Scheuermann.P, "Efficient data reduction with EASE" In. proc.9th international Conference on KDD(2003).
7. Charu.C.Agarwal, Philip S. Yu,"A Survey of synopsis construction in data streams", chapter 9.
8. Chen B, Haas.P, Scheuermann p,"A new two-phase sampling based algorithm for discovering association rules. In Proc. Int conf on ACM SIGKDD (2002).
9. Chuang K,Chen M,Yang .W,"Progresive Sampling for Association Rules based on Sampling Error Estimation", Lecture notes in computer Science,Vol 3518,june 2005 pg 505-515.
10. Cheung,D., Xaio, Y.,Effect of data skewness in parallel mining of association rules, Lecture notes in Computer Science.Volume 1394,Aug 1998,pages 48-60.
11. 11.Das, A., Ng, W.K., and Woon, Y-K.2001.Rapid association rules mining. In Proceedings of the tenth international conference on Information and Knowledge management. ACM press, 474-481.
12. 12.Gibbons, P.B.; and matias, Y.1999.Synopsis data structures for Massive Data sets. In External Memory Algorithms, DIMACS Series in Discrete Mathematics and Theortical Computer Science, 50, 39-70.Providence, Rh.Is. American mathematical Sciences.
13. 13.B.Gu, F.Gu and H.Liu,"Sampling and its application in data mining: A Survey.", Technical report, School of computing, National University of Singapore 2000.
14. 14.Han,J. and Pei.J.2000.Mining frequent patterns by pattern growth: methodology and implications.ACM SIGKDD Explorations Newsletter 2,2,14-20.
15. 15.Hillol Kargupta, Anupam Joshi, Krishnamoorthy Siva Kumar and Yelena Yesha,"Data Mining: Next Generation Challenges and Future Directions", PHI (2005). pp (125-145).
16. D.E.Knuth, The art of computer Programming volume 2, seminumerical Algorithms, Addison Wesley 1981.
17. P.S Levy and S.Zemeshow,"sampling of populations: Methods and Applications", John Wiley and sons, 1991.
18. Li.Y.Gopalan .R,"Effective sampling for mining association rules", Lecture notes in computer Science, volume 3339, Jan 2004 pages 391-401.
19. Manning, A., Keane, J., Data Allocation Algorithm for parallel Association Rule Discovery, Lecture notes in Computer Science, Volume 2035,, Pages 413-420.
20. Mohammed Javeed ,Zaki, Srinivasan Parthasarathy,Wei Li,Mitsunori Ogihara,"Evaluation of Sampling for data mining of Association rules", proc Intn'l workshop research issues in data engineering 1997.
21. Parthasarathy.S," Efficient Progressive Sampling for Association Rules "ICDM 2002:352-361.
22. Parthasarathy, S., Zaki, M.J., Ogihara,M.,Parallel data mining for association rules on shared-memeory systems.Knowledge and Information systems:An International Journal,3(1):1-29,Feb 2001.
23. Sortris Kotsiants, Di mitris Kanellpoulous," Association Rules mining: A Recent Overview" GESTS International Transactions on Computer Science and Engineering, vol 32910, 2006, pp.71-82.
24. Schuster, A. and Wolff, R.(2001), Communication-efficient distributed mining of association rules, in 'Proc.of the 2001 ACM SIGMOD Int'I Conf on management of Data',santa Barbara,California,pp.473-484.
25. Surong Wang, Manoranjan Dash and Ling-Tien Chen,"Efficient Sampling: Application To Image Data, Advances in Knowledge Discovery and DataMining, 9th pacific-asia conference, PAKDD 2005.
26. Tang,.P., Turkia,M.,Parallelizing frequent itemset mining with FP-trees.Technical Report titus.compsci.ualr.edu/~ptang/Ppers/par-fi.pdf,Dept of Computer Science,University of Arkansas at Little Rock,2005.
27. Tien Dung Do, Siu Cheung Hui, Alvis Fong, Mining Frequent Itemsets with Category based Constraints. Lecture notes in Computer Science, Volume 2843, Sep 2003, Pages 76-86.
28. Toivonen .H.(1996),Sampling large databases for association rules in "The VLDB Journal" pp.134-135.
29. Venkatesan T,Vinayaka Pandit,Yogish Sabharwal,"Analysis of sampling techniques for Association rule mining"ACM ,ICDT 2009.
30. J.S. Vitter,"An efficient Algorithm for Sequential random sampling". In ACM Trans Mathematical software, Volume 13910, pages 58-67, Mar 87.
31. Wang,C.,Tjortjis, C.,PRICES: An Efficeint Algorithm for mining Association Rules. Lecture notes in Computer Science, Volume 3177, Jan 2004, and Pages 352-358.

32. Winter, R.; and Auer Bach, K.1998.The Big Time: 1998 Winter Very Large Data Bases Survey. Database Programming Design 11(8).
33. Wojerechowski, M.,Zakrzewiez., M., Dataset Filtering techniques in constraint-based Frequent Pattern Mining,Lecture Notes in Computer Science.Volume 2447,2002,pp.77-83.
34. Wontae Hwang and Dongseung Kim,"Improved Association Rule mining by modified Trimming", Proc of the sixth IEEE international conf on Computer and Information Technology.
35. Yuang, Y., Huang, T.: A matrix algorithm for mining Association Rules.Lecture notes in Computer Science, Volume 3644, Sep 2005, Pages 370-379.
36. Zaki, M J., Parallel and distributed association mining. A survey, IEEE Concurrency,Special Issue on Parallel Mechanisms for Data Mining, 7(4):14—25,December 1999.
37. V.Umarani and M.Punithavalli, "Developing Novel and Effective Approach for Association Rule Mining Using Progressive Sampling", The 2nd International Conference on Computer and Electrical Engineering (ICCEE 2009), Dubai, UAE, December 28-30, 2009. (Accepted for publication).

AUTHORS PROFILE:

Ms. V. Umarani received her Bachelors Degree in Management Sciences and Masters Degree in Computer Application from Bharathiar University. She is currently working as a lecturer and pursuing her doctors' degree in Computer Science in Anna University, Coimbatore, India. Her Area of Interest is Data mining, Data Warehousing and Distributed Databases.

Dr. M.Punithavalli received her Bachelors and Master Degree in Computer Science from Bharathiar University. She received her PhD in Computer Science from Alagappa University. She is currently the Director and Head of Computer Science department, Sri Ramakrishna College of Arts and Science for Women, Coimbatore, India. Her Area of Interest is Data mining, Data Warehousing and Digital Image Processing.