IJDCST @Oct-Nov-2015, Issue- V-3, I-7, SW-10
ISSN-2320-7884 (Online)
ISSN-2321-0257 (Print)

Enhanced Algorithms in Association rule mining

Sampath Kumar Kommineni[1], Dr S. N. Tirumala Rao[2]

[1] Assistant Professor (CSE), Malla Reddy College of Engineering & Technology, Telangana., India.

[2]Professor, Dept. of CSE, Narsaraopeta Engineering College, Narasaraopeta , A.P., India.

Abstract: In data mining, association rule mining is the most important technique in research point of view. Previous data transaction will be analyzed to identify the customer purchasing behavior which is used to improve the business to take the decisions based on the association rules. Researchers presented a lot of approaches and algorithms for determining association rules. This paper discusses few approaches for mining association rules. In this paper we study several aspects in this direction and analyze the previous research. So that we come with the advantages and disadvantages.

Keywords: ARM, Positive Association, Negative Association.

Introduction:

Mining association rules is a vital trip. Past exchange data may be stone-broke all the way down to notice consumer shopping for practices specified the character of business alternative may be improved. The association rules depict the associations among things within the intensive info of consumer transactions. However, the span of the info may be intensive. A huge amount of examination work has been committed to the current region, and led to such methods as k-anonymity [1], data perturbation [2], [3], [4], [5], and data mining based on [6], [7].

Association Rule Mining (ARM) is the most widely used topic in data mining. ARM may be utilised for identifying hidden relationship between things. By given a consumer indicated limit, otherwise referred to as least bolster, the mining of affiliation principles will notice the entire arrangement of incessant examples. That is, before the bottom backing is given, the entire arrangement of normal examples is resolved [8]. Keeping in mind the top goal to recover additional connections among things, shoppers might confirm a moderately bring down least bolster [8].ARM is likewise thought of concerning sector wicker bin examination, that is that the investigation of the itemset which may be stone-broke down when the consumer shopping for within the shopping precinct [8]. it's abundant a similar because the examination of the consumer of getting conduct. Affiliation administers in addition utilised as a section of various zones, for instance, telecommunication systems, sector, hazard administration and stock management etc.[8][9].

In [10] author recommends that data processing is used everywhere the place and plenty of knowledge area unit assembled: in business, to analyses client behavior or optimize production and sales [1].This means the exploration course in an exceedingly few fields. we will utilize ARM and data mining application in social welfare, therapeutic info, arrangement and change of integrity these ways with

alternative approach wide expands the potential conduct and relevancy.

In [11] author proposes that an outsized variety of the specialists area unit for the foremost half targeted around discovering the positive tips simply nonetheless they not find the negative affiliation rules. However, it's in addition essential in examination of perceptive data. It works within the inverse approach of positive principle finding. Be that because it might, issue with the negative affiliation commonplace is it utilizes expansive area and may put aside additional chance to supply the principles as distinction with the customary mining affiliation guideline. So better optimization technique can find a better solution in the above direction.

Related Work:

In 2011, WeiminOuyang et al. [12] suggest three limitations of traditional algorithms for mining association rules.

Firstly, it cannot concern quantitative attributes; secondly, it finds out frequent itemsets based on the single one userspecified minimum support threshold, which implicitly assumes that all items in the data have similar frequency; thirdly, only the direct association rules are discovered. They propose mining fuzzy association rules to address the first limitation. In this they put forward a discovery algorithm for mining both direct and indirect fuzzy association rules with multiple minimum supports to resolve these three limitations.

In 2012, YihuaZhong et al. [13] suggest that association rule is an important model in data mining. However, traditional association rules are mostly based on the support and confidence metrics, and most algorithms and researches assumed that each attribute in the database is equal. In fact, because the user preference to the item is different, the mining rules using the existing algorithms are not always appropriate to users. By introducing the concept of weighted dual confidence, a new algorithm which can mine effective weighted rules is proposed by the authors. The case studies show that the algorithm can reduce the large number of meaningless association rules and mine interesting negative association rules in real life.

In 2012, He Jiang et al. [14] support the technique that allows the users to specify multiple minimum supports to reflect the natures of the itemsets and their varied frequencies in the database. It is very effective for large databases to use algorithm of association rules based on multiple supports. The existing algorithms are mostly mining positive and negative association rules from frequent itemsets. But the negative association rules from infrequent itemsets are ignored. Furthermore, they set different weighted values for items according to the importance of each item. Based on the above three factors, an algorithm for mining weighted negative association rules from infrequentitemsets based on multiple supports(WNAIIMS) is proposed by the author.

In 2012, IdhebaMohamad Ali O. Swesi et al. [15] study is to develop a new model for mining interesting negative and positive association rules out of a transactional data set. Their proposed model is

integration between two algorithms, the Positive Negative Association Rule (PNAR) algorithm and the Interesting Multiple Level Minimum Supports (IMLMS) algorithm, to propose a new approach (PNAR_IMLMS) for mining both negative and positive association rules from the interesting frequent and infrequent item sets mined by the IMLMS model. The experimental results show that the PNAR_IMLMS model provides significantly better results than the previous model.

In 2012, WeiminOuyang [16] suggest that traditional algorithms for mining association rules are built on the binary attributes databases, which has three limitations. Firstly, it cannot concern quantitative attributes; secondly, only the positive association rules are discovered; thirdly, it treat each item with the same frequency although different item may have different frequency. So he puts forward a discovery algorithm for mining positive and negative fuzzy association rules to resolve these three limitations.

In 2012, XiaofengZheng et al. [17] presented the theory, question and resolution of application of rough set in mining association rules. And it presented resolve the relation of support, confidence and the amount of rules by rough set analysis originally. According to the authors the entire conclusions were proved in data mining in provincial road transportation management information System.

Association Rules:

From 1993 [18] the task of association rule mining has received a great deal of attention. Today the mining of such rules is still one of the most popular pattern discovery methods in KDD. In brief, an association rule is an expression $X \Rightarrow Y$, where X and Y are sets of items. The meaning of such rules is quite intuitive: Given a database D of transactions where each transaction $T \epsilon D$ is a set of items - $X \Rightarrow Y$ expresses that whenever a transaction T contains X than T probably contains Y also. The probability or rule confidence is defined as the percentage of transactions containing Y in addition to X with regard to the overall number of transactions containing X. That is, the rule confidence be understood as the conditional probability p $(Y \subseteq T | X \subseteq T)$.

The idea of mining association rules originates from the analysis of market-basket data where rules like A customer who buys products x1 and x2 will also buy product y with probability are found. Their direct applicability to business problems together with their inherent understandability even for non data mining experts made association rules a popular mining method. Moreover it became clear that association rules are not restricted to dependency analysis in the context of retail applications, but are successfully applicable to a wide range of business problems

When mining association rules there are mainly two problems to deal with: First of all there is the algorithmic complexity. The number of rules grows exponentially with the number of items. Fortunately today's algorithms are able to efficiently prune this immense search space based on minimal thresholds for quality measures on the rules.

Second, interesting rules must be picked from the set of generated rules. This might be quite costly because

the generated rule sets normally are quite large { e.g. more than 100; 000 rules are not uncommon } and in contrast the percentage of useful rules is typically only a very small fraction. The work on concerning the second problem mainly focuses on supporting the user when browsing the rule set, e.g. [19] and the development of further useful quality measures on the rules, e.g. [20; 21; 22].

Algorithms:

We first give an overview of the AIS [18] and SETM [23] algorithms against which we compare the performance of the Apriori and AprioriTid algorithms. We then describe the synthetic datasets used in the performance evaluation and show the performance results. Finally, we describe how the best performance features of Apriori and AprioriTid can be combined into an AprioriHybrid algorithm and demonstrate its scale-up properties.

The most popular algorithm of this type is Apriori were also the downward closure property of itemset support was introduced. Apriori makes additional use of this property by pruning those candidates that have an infrequent subset before counting their supports. This optimization becomes possible because BFS ensures that the support values of all subsets of a c candidate are known in advance. Apriori counts all candidates of a cardinality k together in one s an over the database. The critical part is looking up the candidates in each of the transactions. The items in each transaction are used to des end in the hashtree. Whenever we reach one of its leafs, we find a set of candidates having a common prefix that is contained in the transaction. Then these candidates are searched in the transaction that has been en coded as a bitmap before. In the case of success the counter of the candidate in the tree is incremented. AprioriTID is an extension of the basi Apriori approach. Instead of relying on the raw database AprioriTID internally represents each transaction by the current candidates it contains. With AprioriHybrid both approaches are combined. To some extent also SETM [23] is an Apriori (TID)-like algorithm which is intended to be implemented directly in SQL.

The AIS Algorithm

Candidate itemsets are generated and counted on-the-y as the database is scanned. After reading a transaction, it is determined which of the itemsets that were found to be large in the previous pass are contained in this transaction. New candidate itemsets are generated by extending these large itemsets with other items in the transaction. A large itemset l is extended with only those items that are large and occur later in the lexicographic ordering of items than any of the items in l. The candidates generated from a transaction are added to the set of candidate itemsets maintained for the pass, or the counts of the corresponding entries are increased if they were created by an earlier transaction.

The SETM Algorithm

The SETM algorithm [13] was motivated by the desire to use SQL to compute large itemsets. Like AIS, the SETM algorithm also generates candidates on- the-y based on transactions read from the database. It thus generates and counts every

candidate itemset that the AIS algorithm generates. However, to use the standard SQL join operation for candidate generation, SETM separates candidate generation from counting. It saves a copy of the candidate itemset together with the TID of the generating transaction in a sequential structure. At the end of the pass, the support count of candidate itemsets is determined by sorting and aggregating this sequential structure.

SETM remembers the TIDs of the generating transactions with the candidate itemsets. To avoid needing a subset operation, it uses this information to determine the large itemsets contained in the transaction read. Lk _ Ck and is obtained by deleting those candidates that do not have minimum support. Assuming that the database is sorted in TID order, SETM can easily _nd the large itemsets contained in a transaction in the next pass by sorting Lk on TID. In fact, it needs to visit every member of Lk only once in the TID order, and the candidate generation can be performed using the relational merge-join operation [13].

The disadvantage of this approach is mainly due to the size of candidate sets Ck. For each candidate itemset, the candidate set now has as many entries as the number of transactions in which the candidate itemset is present. Moreover, when we are ready to count the support for candidate itemsets at the end of the pass, Ck is in the wrong order and needs to be sorted on itemsets. After counting and pruning out small candidate itemsets that do not have minimum support, the resulting set Lk needs another sort on TID before it can be used for generating candidates in the next pass.

Experiments:

AprioriHybrid scales up as the number of transactions is increased from 100,000 to 10 million transactions. We used the combinations (T7.I3), (T11.I5), and (T21.I7) for the average sizes of transactions and itemsets respectively. The sizes of these datasets for 10 million transactions were 242MB, 442MB and 842MB respectively. The minimum support level was set to 0.75%. The execution times are normalized with respect to the times for the 100,000 transaction datasets in the first graph and with respect to the 1 million transaction dataset in the second. As shown, the execution times scale quite linearly.

Next, we examined how AprioriHybrid scaled up with the number of items. We increased the number of items from 1000 to 10,000 for the three parameter settings T5.I2.D100K, T10.I4.D100K and T20.I6.D100K. All other parameters were the same as for the data in Table 3. We ran experiments for a minimum support at 0.75%, and obtained the results shown in Figure 9. The execution times decreased a little since the average support for an item decreased as we increased the number of items. This resulted in fewer large itemsets and, hence, faster execution times.

Finally, we investigated the scale-up as we increased the average transaction size. The aim of this experiment was to see how our data structures scaled with the transaction size, independent of other factors like the physical database size and the number of large itemsets. We kept the physical size of the database roughly constant by keeping the product of the average transaction size and the number of transactions constant. The number of transactions

ranged from 200,000 for the database with an average transaction size of 5 to 20,000 for the database with an average transaction size 50. Fixing the minimum support as a percentage would have led to large increases in the number of large itemsets as the transaction size increased, since the probability of itemset being present in a transaction is roughly proportional to the transaction size. We therefore fixed the minimum support level in terms of the number of transactions. The results are shown in Figure 1.

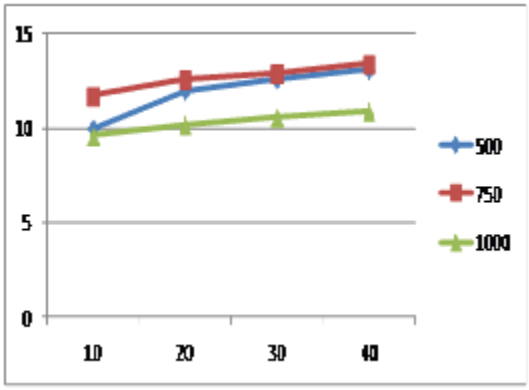

Figure 1, Transaction Size

The numbers in the key (e.g. 500) refer to this minimum support. As shown, the execution times increase with the transaction size, but only gradually. The main reason for the increase was that in spite of setting the minimum support in terms of the number of transactions, the number of large
itemsets increased with increasing transaction length. A secondary reason was that finding the candidates present in a transaction took a little longer time.

Conclusion:

In this paper, we presented two new algorithms, Apriori and AprioriTid, for discovering all significant association rules between items in a large database of transactions. We compared these algorithms to the previously known algorithms, the AIS and SETM algorithms. We presented experimental results, showing that the proposed algorithms always outperform AIS and SETM. The performance gap increased with the problem size, and ranged from a factor of three for small problems to more than an order of magnitude for large problems.

References:

[1] Sweeney, L. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5):557-570, 2002.

[2] Agrawal, R. and Srikand R. Privacy preserving data mining. In Proc. Of ACM SIGMOD Conference, pp. 439-450, 2000.

[3] Chen, K. and Liu. L. A random rotation perturbation approach to privacy data classification. In Proc of IEEE Intl. Conf. on Data Mining (ICDM), pp. 589-592, 2005.

[4] Xu, S., Zhang, J., Han, D. and Wang J. Singular value decomposition based data distortion strategy for privacy distortion. Knowledge and Information System, 10(3):383-397, 2006.

[5] Mukherjeee, S., Chen, Z. and Gangopadhyay, A. A privacy-preserving technique for Euclidean distance-based mining algorithms using Fourier related transforms. Journal of VLDB, 15(4):293-315, 2006.

[6] Vaidya, J. and Clifton, C. Privacy preserving k-means clustering over vertically partitioned data. In Prof. of ACM SIGKDD Conference, pp.206-215, 2003.

[7] Vaidya, J., Yu, H. and Jiang, X. Privacy preserving SVM classification. Knowledge and Information Systems, 14:161-178, 2007.

[8] Ms. KumudbalaSaxena, Dr. C.S. Satsangi, "A Non Candidate Subset-Superset Dynamic Minimum Support Approach for sequential pattern Mining", International Journal of Advanced Computer Research (IJACR),Volume-2 Number-4 Issue-6 December-2012.

[9] Dr. Manish Shrivastava, Mr. Kapil Sharma, MR. Angad Singh, "Web Log Mining using Improved Version of Proposed Algorithm", International Journal of Advanced Computer Research (IJACR),Volume 1 Number 2 December 2011.

[10] PragatiShrivastava, Hitesh Gupta," A Review of Density-Based clustering in Spatial Data", International Journal of Advanced Computer Research (IJACR), Volume-2 Number-3 Issue-5 September-2012.

[11] Nikhil Jain,VishalSharma,MaheshMalviya, "Reduction of Negative and Positive Association Rule Mining and Maintain Superiority of Rule Using Modified Genetic Algorithm", International Journal of Advanced Computer Research (IJACR) ,Volume-2 Number-4 Issue-6 December-2012.

[12] WeiminOuyang; Qinhua Huang, "Mining direct and indirect fuzzy association rules with multiple minimum supports in large transaction databases," Fuzzy Systems and Knowledge Discovery (FSKD), Eighth International Conference on , vol.2, no., pp.947,951, 26-28 July 2011.

[13] YihuaZhong; Yuxin Liao, "Research of Mining Effective and Weighted Association Rules Based on Dual Confidence," Computational and Information Sciences (ICCIS), Fourth International Conference on , vol., no., pp.1228,1231, 17-19 Aug. 2012.

[14] He Jiang, Xiumei Luan and Xiangjun Dong," Mining Weighted Negative Association Rules from Infrequent Itemsets Based on Multiple Supports", International Conference on Industrial Control and Electronics Engineering, 2012.

[15] IdhebaMohamad Ali O. Swesi, Azuraliza Abu Bakar, AnisSuhailis Abdul Kadir," Mining Positive and Negative Association Rules from Interesting Frequent and Infrequent Itemsets", 9th International Conference on Fuzzy Systems and Knowledge Discovery, 2012.

[16] WeiminOuyang," Mining Positive and Negative Fuzzy Association Rules with Multiple Minimum Supports", International Conference on Systems and Informatics, 2012.

[17] XiaofengZheng and JianminXu," Studies on the Application of Rough set Analysis in Mining of Association Rules and the Realization in Provincial Road Transportation Management Information System", International Conference on Industrial Control and Electronics Engineering,2012.

[18] R. Agrawal, T. Imielinski, and A. Swami. Mining asso- iation rules between sets of items in large databases. In Pro . of the ACM SIGMOD Int'l Conf. on Management of Data (ACM SIGMOD '93), Washington, USA, May 1993.

[19] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkamo. Finding interesting rules from large sets of dis overed asso iation rules. In Pro . of the 3rd Int'l Conf. on Information and

Know ledge Management, Gaithersburg, Maryland, 29. Nov - 2. Dec 1994.

[20] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynami itemset ounting and impli ation rules for market basket data. In Pro . of the ACM SIGMOD Int'l Conf. on Management of Data, 1997.

[21] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing asso iation rules to orrelations. In Pro . of the ACM SIGMOD Int'l Conf. on Management of Data (ACM SIGMOD '97), 1997.

[22] C. Silverstein, S. Brin, R. Motwani, and J. D. Ullman. S alable te hniques for mining ausal stru tures. In Pro . of 1998 ACM SIGMOD Int'l Conf. on Management of Data, Seattle, Washington, USA, June 1998.

[23] M. Houtsma and A. Swami. Set-oriented mining of association rules. Research Report RJ 9567, IBM Almaden Research Center, San Jose, California, October 1993.