DMSA TECHNIQUE FOR FINDING SIGNIFICANT PATTERNS IN LARGE DATABASE

Saravanan.Suba

Assistant Professor of Computer Science
Kamarajar Government Art & Science College
Surandai, TN, India-627859
Email:saravanansuba@rediffmail.com

Dr. Christopher.T

Assistant Professor Of Computer Science
Government Arts and Science College
Coimbatore ,TN, India-641018
Email:Chris.hodcs@gmail.com

Abstract

Frequent pattern mining in databases plays a vital role in many data mining tasks like classification, sequential patters, clustering, association rules analysis etc. There are numerous mining algorithms for finding association rules. One of the most common algorithms is Apriori. It is used to mine frequent item sets from large database. It uses the support statistical measure for pruning the frequent items. When the higher minimum support is used to reduce the number of frequent items for rules generation, this algorithm misses some of the bigger combination of significant frequent items. This proposed DMSA (Dynamic Minimum Support Apriori) generates significant frequent items in large database for association rules generation.

Keywords: Apriori; Frequent itemset; Higher minimum support; DMSA

I. INTRODUCTION

The data mining refers to extract or mine knowledge from huge volume of data [1]. Data mining plays an essential role in numerous applications like market–basket analysis, fraud detection … etc. One of the best popular descriptive data mining methods is association rule mining [2]. Since its introduction [3], Association rule mining has become one of the essential data mining tasks and has involved tremendous interest among data mining researches and professionals [4]. It is a well-known method for discovering correlations between variables in large databases. Consider the example of shopping mall. The transactional database of shopping mall consists of two attributes, transaction id, and items purchased by the consumer. Each Transaction id is distinctive. The mined patterns are set of items that are most frequent in database.

For example, it has to find out how many of consumers purchase milk and sugar together. And how many consumers purchase bread and jam together. To find out such facts apply market basket analysis and discover pattern. Domain expert can use this detail for identifying the consumer purchasing behaviours to maximize the profit of the organization. So frequent pattern mining is most powerful problem in association rule mining.

Many of the algorithms are based on traditional algorithm of association rule mining [3][5]. There are two key functions in Apriori to discover association rules. Firstly, it discovers frequent item set based on minimum support count. After that, minimum confidence is used to discover association rules between frequent items. Two statistical measures that control the activity of association rule mining are support and confidence [6]. For an association rule $X \rightarrow Y$ and total number of transactions is denoted as N, the support and confidence can be mathematically represented as follows

$$\text{Support } (X \rightarrow Y) = \sum(XUY)/N \text{ and}$$

$$\text{Confidence } (X \rightarrow Y) = \sum (XUY)/ \sum X.$$

Most of the earlier studies implement Apriori-like algorithms, which improves algorithm strategy and structure. This paper introduces the DMSA techniques to find significant frequent item sets.

The rest of the paper is organized as follows: Related works and background are described in section 2. The proposed algorithm is discussed in section 3. Experimental results and discussion are given in section 4. The conclusion and the ideas for future work are written in section 5.

II. RELATED WORKS

Mining of frequent itemsets is an essential phase in association mining which finds frequent itemsets in historical database. Apriori algorithm is used to discover frequent itemsets in large dataset and it was introduced by Agrawal et al. [3]. Numerous modifications on Apriori algorithm were dedicated [5][7][8][9][10][11][12][13][14][15][16][17][18][19] and they used single minimum support for extracting frequent pattern mining. Some of the algorithms [20][21][22] were presented with multiple minimum support to improve the generation of rare items combination. The following proposed method will discover both normal as well as rare bigger combination frequent items using Dynamic Minimum Support.

III. THE DSMA

1. **Algorithm:** DMSA
2. **Input:** Dataset D, initial minimum support value σ, number of items I, no of transactions in data set N
3. **Output:** frequent item set F
4. **Begin**
5. Assign all items as candidate(c_1-1-item candidate set);
6. DMS← σ;
7. DMS_INT ←DMS/(I-1);
8. **Repeat**
9. {
10. Scan the dataset D and count the occurrences of each candidate and record it (L_i - candidate with count);
11. Take each candidate with count from C_i and check if candidate support value >=DMS then
12. {
13. add candidate to F;
14. }
15. DMS←DMS-DMS_INT;
16. Generate next candidate set $C_{i+1;}$
17. } **Until** C_{i+1}=null
18. End.

Let first consider an example of five transactions, six items and initial support value of 50%. The transactions are given in Table1.

Table 1: Transactions

Trans id	items
100	I1,I2,I3,I4
200	I1,I2,I4
300	I1,I5,I6
400	I1,I4,I5
500	I2,I4,I5

Firstly, scan all transactions to find frequent 1-itemset $L_{1,}$ which contains the items and their support counts. Then set initial support DMS value as 50 and calculate DMS interval using DMS/ (I-1). The frequent 1-item set is shown in Table 2.

Table 2: Frequent 1-itemsets

Items	Counts	% Of Support
I1	4	80%
I2	3	60%
I3	1	20%
I4	4	80%
I5	3	60%
I6	1	20%

The items I3 and I6 are not satisfied with DMS value. So they are deleted from the process. The next step is to generate candidate 2-itemset from L_1. It is shown in *Table 3.*

Table: 3 frequent 2-itemset

Items	Count	% of Support
I1,I2	2	40%
I1,I4	3	60%
I1,I5	2	40%
I2,I4	3	60%
I2,I5	1	20%
I4,I5	2	40%

Now update the DMS value as DMS-DMS_INT (i.e. DMS=40) and find the L_2. From the above Table 3 the candidate (I2, I5) is not satisfied with DMS value. So it is deleted from the process. This process will be continued until the candidate set becomes empty.

IV. EXPERIMENTAL RESULTS

The intel® core ™ i5-2450m CPU @2.5 GHZ,4.0GB RAM ,64bit windows 7 operating system and NetBeans IDE 8.0.2 were used to conduct the experiment. The synthetic Data set of 100,200, 400 and 1000 with 10 items were created to compare the proposed DMSA with Apriori in terms of number of significant frequent items finding.

The Table 4 shows the number of significant frequent items generated by DMSA and Apriori with 20% minimum support.

Table 4: Number of Frequent Items

Number Of Transactions	Apriori	DMSA
100	119	356
200	109	287
400	79	159
1000	84	196

When 100 transactions are used, Apriori generates 119 frequent items where as DMSA generates 356 frequent items. Likewise, the DMSA generate more significant frequent items than Apriori. The following figure-1 demonstrates the number of frequent items generated by Apriori and DMSA for given number of transactions.

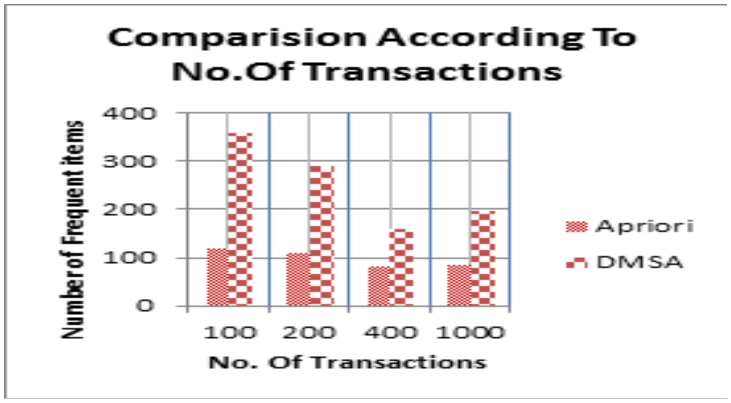

Figure 1: Number of frequent item generation

It is observed that the DMSA generates more significant frequent items (including significant rare bigger combination frequent items which may be more profitable) than Apriori . The domain expert should identify the rare bigger combination significant frequent items generated by DMSA which gives more profit. The organization should use the details given by the domain expert to increase the profit of the organization.

V. CONCLUSION

If the support value is less, the Apriori generate lot of frequent item sets which are very difficult to filter and find significant frequent items. If the support values is high, Apriori generate very less frequent item sets .So this may miss some of the significant frequent item sets. The DMSA algorithm generates significant frequent item sets(including rare bigger combination frequent items) for finding association rules. The domain expert can identify some rare bigger combination frequent items which gives more profit from frequent items generated by DMSA. In future DMS should be applied on various real time data sets and modified Apriori algorithms.

REFERENCES

[1] G.K.Gupta, "Introduction to Data Mining with Case Studies", PHI Learning private limited, New Delhi ,2009.

[2] S.Shankar and T.Purusothaman, "Utility Sentient Frequent Item Set Mining And Association Rule Mining: A Literature Survey And Comparative Study", International Journal of Soft Computing Applications ISSN: 1453-2277 Issue 4, pp.81-95,2009.

[3] Agrawal, R., Imielinski, T., and Swami, A. N, "Mining Association Rules Between Sets Of Items In Large Databases", In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-216,1993.

[4] M. J. Zaki and C.J. Hsiao, "CHARM: An Efficient Algorithm For Closed Association Rule Mining", Technical Report 99-10, Computer Science Dept., Rensselaer Polytechnic Institute, October 1999.

[5] Agrawal, R. and Srikant, R, "Fast Algorithms For Mining Association Rules", In Proc. 20th Int. Conf. Very Large Data Bases, 487-499, 1994.

[6] Saravanan.Suba and Dr. Christopher., "A Study On Milestones Of Association Rule Mining Algorithms In Large Databases", International Journal of Computer Applications (0975 – 888) Volume 47– No.3. T , June2012

[7] Sotiris Kotsiantis and Dimitris Kanellopoulos, "Association Rules Mining: A Recent Overview", GESTS International Transactions on Computer Science and Engineering, Vol.32, No: 1, pp. 71-82, 2006.

[8] Anurag Choubey, Ravindra Patel.J, L. Rana, "A Survey Of Efficient Algorithms And New Approach For Fast Discovery Of Frequent Item Set For Association Rule Mining", International Journal of Soft Computing and Engineering, May 2011.

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

[10] 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.

[11] 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.

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

[13] C. Hidber, "Online Association Rule Mining", In A. Delis, C. Faloutsos, and S.Ghandeharizadeh, editors, Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data, volume 28(2) of SIGMOD Record, pp. 145–156, ACM Press, 1999.

[14] V.Umarani et al, "A Study on Effective Mining of Association Rules from Huge Databases", International journal of computer science and research,Vol .1,issue , 2010.

[15] Toivonen H, "Sampling Large Databases For Association Rules ", In VLDB Journal, pp. 134-145, 1996.

[16] Kun-Ta Chuang, Ming-Syan Chen, Wen-Chieh Yang, "Progressive Sampling for Association Rules Based on Sampling Error Estimation", Lecture Notes in Computer Science, Volume 3518, pp. 505 – 515, 2005.

[17] V.Umarani and M.Punithavalli, "On Developing an Effectual Progressive Sampling Based Approach for Association Rule Discovery", In the proceedings of 2nd IEEE International Conference on Information and data Engineering (2nd IEEE ICIME 2010), Chengdu ,China, April 2010 .

[18] Savesere A, Omiecinski E, and Navathe S, "An Efficient Algorithm For Mining Association Rules In Large Databases", In Proceedings of 20th International Conference on VLDB, 1995.

[19] Parthasarathy.S, Zaki, M.J.J and Ogihara M "Parallel Data Mining For Association Rules On Shared- Memory Systems", Knowledge and Information Systems: An International Journal, 3(1), pp.1-29 , 2001.

[20] Lin, Bing, Hsu, wynne and Ma, yiming, "Mining Association Rules with Multiple Minimum Support" ,KDD99, 1999.

[21] Ya-Han Hu, Yen-Liang Chen, " Mining Association Rules With Multiple Minimum Support : A New Mining Algorithm And A Support Tuning Mechanism " Elsevier, 2004.

[22] R.Uday Kiran ,P.krishna Reddy" An Improved Multiple Minimum Support based Approach to Mine Rare Association Rules", IEEE Symposium on computational Intelligence and Data Mining,2009.