International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Improvement in Apriori Algorithm with New Parameters

Reeti Trikha[1], Jasmeet Singh[2]

[1]Computer Science & Engineering, RIMT-IET, Mandi Gobindgarh, Punjab, India
Affiliated to Punjab Technical University, Jalandhar, Punjab, India

[2]Computer Science & Engineering, RIMT-IET, Mandi Gobindgarh, Punjab, India
Affiliated to Punjab Technical University, Jalandhar, Punjab, India

Abstract: ***Data Mining or Knowledge Discovery in Databsaes is an advanced approach which refers to the extraction of previously unknown and useful information from large databases. Association Rule Mining is an important technique of data mining. This technique emphasis on finding interesting relationships. For understanding these relationships, a technique called Market Basket Analysis has been introduced in Data Mining. This helps in understanding the customer behaviour more easily so that frequent patterns can be generated. Apriori algorithm is used in association rule mining for generating frequent patterns. But it generates patterns only on the basis of presence and absence of items, resulting into lack of efficient results. So new parameters have been included in this paper which will be helpful in giving maximum profit to the business organizations. This paper shows that how addition of new parameters improve the efficiency of Apriori algorithm by comparing the results of improved algorithm with the results of traditional Apriori algorithm.***

Keywords: Data Mining, KDD, Association Rule Mining, Apriori, Market Basket Analysis, Support, Confidence, Profit, Weight, Q-factor, PW-Factor.

1. Introduction

1.1 Data Mining

Data mining [11], also popularly known as Knowledge Discovery in Databases (KDD), refers to the non-trivial extraction of implicit, previously unknown and potentially useful information from data in large databases. Generally, both data mining and knowledge discovery in databases are treated as synonyms but the fact is that data mining is actually a part of knowledge discovery process. Moreover, the patterns discovered depends completely on the technique which we employ. The patterns vary from one technique to another technique. Its main objectives are predicting and describing the data. The following steps are involved in the KDD process as shown in fig 1.1:

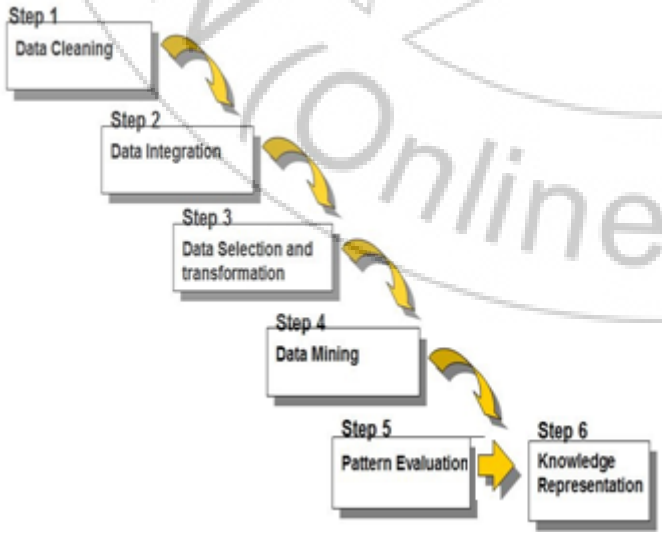

Figure 1.1: KDD Process

Mainly it involves three processes as:

1) **Pre-processing:** This process is executed before data mining techniques are actually applied on the data. It involves data cleansing, data integration, data selection and transformation.
2) **Data Mining:** It is the main step of KDD process. It uses different algorithms to extract the hidden information.
3) **Post Processing:** It's the last step of KDD process which evaluates the mining result according to requirement of user and domain knowledge. It presents knowledge only if results are satisfactory, otherwise more processes are run till desired results are achieved.

Data mining involves many useful techniques. Association Rule Mining is one among them. Its main job is to find interesting relationships or association rules between databases which helps in various data mining projects.

1.2 Association Rule Mining(ARM)

1.2.1 Introduction

Association Rule Mining [4] is a widely-used approach in data mining. It is used for generating association rules which help in revealing interesting relationships in large databases. These association rules not only help in describing relationships in large databases but also used for discriminating between different kinds of databases. Association rule mining deals with not only finding relationships but with interesting relationships.

1.2.2 Concepts Of Association Rule Mining:

Suppose we have I as a set of items, D as a set of transactions, then association rule is an implication of the form X ->Y, where X, Y are subsets of I, and X, Y do not intersect.

Every rule involves mainly two measures- support and confidence.

- *Support:* It's the *probability that* both X and Y will come together in a transaction.
- Support(X, Y) = n (XUY)/N
- N=Total no .of transactions.
- *Confidence:* It's the probability that follows a condition that is a transaction having X also contains Y.
- Confidence(X, Y) = support (XUY)/support(X)
- *Itemset:* It is a set of items in a transactional database.
- *k-itemset:* It is a set of k-items in a transactional database.

1.2.3 Market Basket Analysis

Market Basket Analysis [9] is of great help in association rule mining. It's of great help in studying the behaviour of the customers. In this, the items to be purchased by the customer are placed together and the purchasing behaviour of customers is analysed. In this way the organizations get an idea that which items to be selected more and how much quantity of those items needs to be raised for gaining maximum profit. So they place the items accordingly. Moreover, in this way the organizations get an idea about customer's requirements easily.

2. Apriori Algorithm

The Apriori algorithm [10] was proposed by R. Agrawal and R. Srikant in 1994. They introduced an innovative approach of finding association rules on large scale. With the help of this algorithm frequent itemsets are generated from transactional database. It generates these itemsets based upon the minimum support threshold.

2.1 Basic Concepts Of Apriori Algorithm:

Some basic terms related to Apriori are:

- *Itemset*: It's a collection of items in a database.
- *Transaction*: It's a database entry which contains a collection of items.
- *Support:* Interesting association rule can be measured with the help of support criteria. Support is nothing but how many transactions have such itemsets that match both sides of the
- Implications in the association rule.
- Support (i) =Count (i)/total transaction
- *Candidate Itemset (Li):* Items which are only used for the processing. Candidate itemsets are all possible combination of itemsets.
- *Minimum Support:* It's a condition which helps to eliminate the non-frequent items from database.
- *Frequent Itemset (Large Itemset (Li)):* The itemsets which satisfies the minimum support criteria are known as frequent itemsets.

2.2 Working of Apriori Algorithm:

Apriori algorithm is applied on the transactional databse. It generates both frequent and non-frequent itemsets [3] as a result. But the frequent itemsets are further used whereas non-frequent itemsets are discarded. This process continues till we end with frequent itemsets only. Basically this process can be summarized in two basic concepts as:

a) Joining: In this step candidate itemsets are joined.
b) Pruning: In this step frequent itemsets are discovered and used whereas non-frequent itemsets are discarded.

Apriori Algorithm Steps:

1) First, the set of frequent 1-itemsets is found. (Known as C1).
2) Then support is calculated by counting the occurrence of the item in transactional database.
3) After that we will prune the C1 using minimum support Criteria. The item which satisfies the minimum support criteria is taken into consideration for the next process and which is known as L1.
4) Then again candidate set generation is carried out and the 2-itemset which is generated known as C2.
5) Again we will calculate the support of the 2- Itemset.and we will prune C2 using minimum support and generate L2.
6) This process continues till there is no Candidate set and frequent itemsets can be generated.

3. Problem Statement and Methodology

3.1 Enhancement of Attributes of Apriori Algorithm in Association Rule Learning

Proposed Research work is based on the improvement of algorithm which is based upon the Apriori algorithm that will enhance the efficiency by making a model of prototype which will be beneficial in overcoming the shortcomings of Apriori algorithm. The Apriori algorithm is theoretically and experimentally analyzed which is the most established algorithm for frequent itemset mining. The work is focused on Apriori algorithm. The proposed improvement of algorithm is implemented in C and the work also involves the use of some tools of data mining.

3.2 Primary Objective

The work is the efficient frequent patterns extracted through utilization of attributes in order to overcome the disadvantages of association rule mining and Apriori algorithm. It has the following main objectives:

- To identify interesting patterns from transactional database.
- To improve efficiency of Apriori algorithm and association rule mining by generating interesting patterns using attributes, i.e. profit ratio using Q-factor and the Profit-Weighing factor.
- The interesting patterns should give maximum profit to the business.

3.3 Steps of Proposed Methodology

Following are the steps involved in the proposed methodology. It will tell that how the proposed work has been done. Firstly, a transactional database is assumed on which the proposed methodology is to be applied.

STEP 1: Firstly, a database is assumed which consists of number of items to be purchased by the customer and total profit achieved by the items. Profit ratio for each item is calculated by applying Q-Factor.
STEP 2: Now, a transactional database is assumed which consists of number of items that are purchased by the customer and total number of transactions in which the customer purchases the items.
STEP 3: Now, the Apriori algorithm of association rule mining is applied in order to determine the frequency of each itemset.
STEP 4: The confidence measure of each itemset is calculated.
STEP 5: The itemsets are sorted based upon the user specified minimum confidence.
STEP 6: Now, the Profit-Weighing Factor is applied on the sorted itemsets.
STEP 7: Output is the frequent itemsets which are giving maximum profit to the business.

4. Results and Discussion

The various results that are obtained by the implementation of the proposed improved algorithm are discussed below. The Apriori algorithm is applied to the dataset containing 5 transactions and 3 itemsets. This Apriori algorithm is applied using the data mining tool - Tanagra. The various parameters for Apriori algorithm are entered i.e. the values for minimum support and minimum confidence is given by the end user.

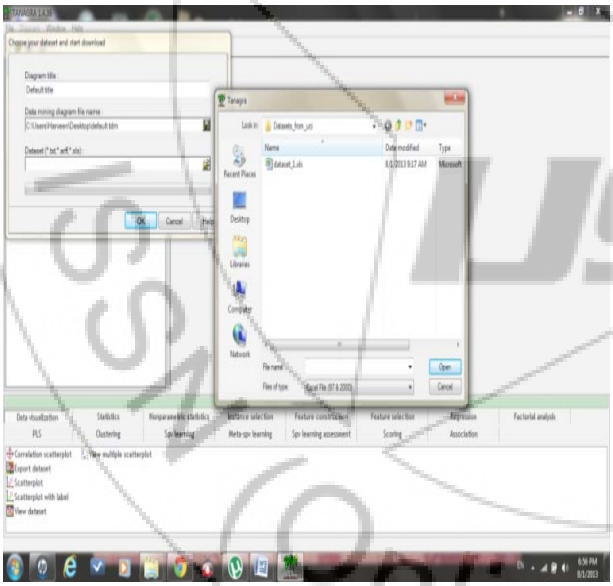

Figure 4.1: Importing the dataset in Tanagra

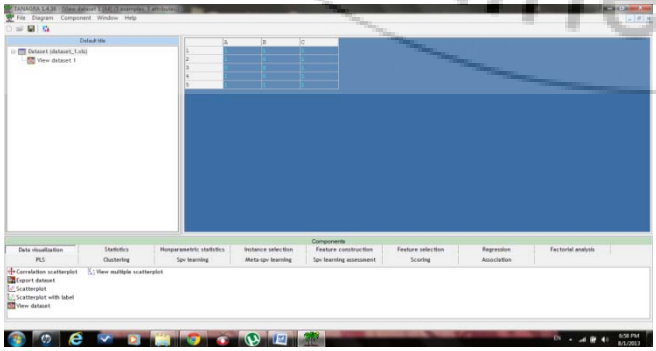

Figure 4.2: Viewing the dataset in Tanagra

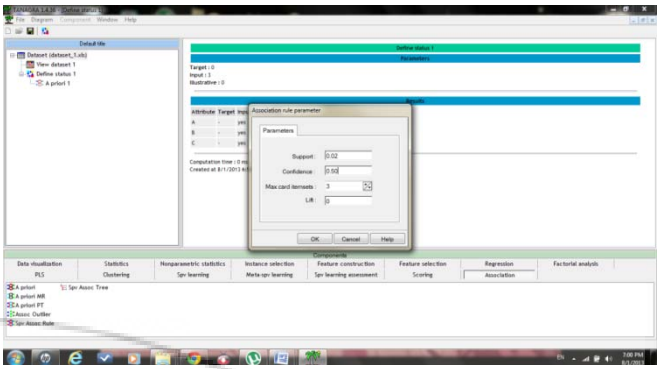

Figure 4.3: Specifying the values of parameters

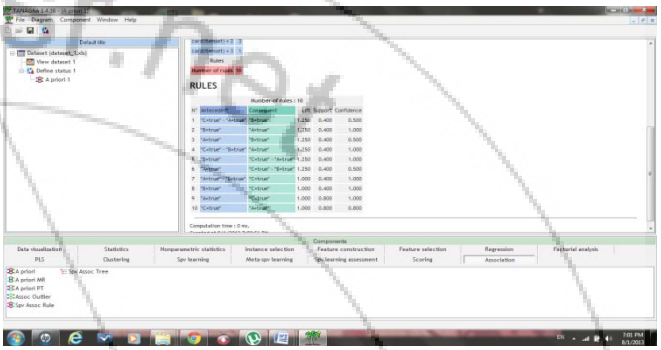

Figure 4.4: Rules obtained using Apriori algorithm in tanagra

The various frequent itemsets obtained with the help of above method are ab, ac, bc, abc. The above algorithm takes into consideration only the presence and absence of an item in the transactional database and do not consider the profit associated with each item. The proposed improved algorithm considers the profit attribute of each item and thus calculates the profit ratio using Q-factor for each item. This is beneficial as it gives the total amount of profit an itemset is giving to the seller. The proposed improved algorithm is applied to the same dataset. Here dataset containing 5 transactions and 3 items is considered whose implementation output is shown below. The whole implementation is done in multiple steps using C Language.

```
Enter dataset name =XYZ
 No. of Transactions: 5
 No. of Items: 3
1    1    0    1    1
1    0    0    0    1
1    1    1    1    1
STEP 1:Transactions :5
          Items :3
STEP 2: Entering Actual Profit For Items
          Enter Actual Profit For Item a:90
          Enter Actual Profit For Item b:10
          Enter Actual Profit For Item c:60
          Total Sum of Profit of the items:160
```

STEP 3: Given Transactional Database containing existing values in database for items
Existing Values in Database for Item a: 1 1 0 1 1
Existing Values in Database for Item b: 1 0 0 0 1
Existing Values in Database for Item c: 1 1 1 1 1
STEP 4: Calculation of Item Profit Ratio Using Q FACTOR:
Profit Ratio using Q-FACTOR for ITEM a: 0.562500
Profit Ratio using Q-FACTOR for ITEM b: 0.062500
Profit Ratio using Q-FACTOR for ITEM c: 0.375000
STEP 5: Calculation of Item Frequency using SUPPORT
Enter minimum support value:2
Frequency of Two-Itemsets:
ac: 4
bc: 2
ab: 2
ac: 4
bc: 2
Frequency of Three-Itemsets will be:
abc:2
STEP 6: Calculation of CONFIDENCE measure
Enter minimum confidence value :50
Sorted itemsets values based on confidence:
ab:50.000000
abc: 50.000000
ac:100.000000
bc: 100.000000
STEP 7: Calculation of P-W factor for Itemsets
Enter Profit Threshold value for Itemsets :1.0
FINAL OUTPUT:P-W value of Profitable Itemsets:
ac:3.750000
abc:2.00000
ab:1.250000

The above results can be shown with the help of graph which clearly explains that the frequent itemset which is giving 100% confidence with the classical Apriori algorithm may not provide high profit to the seller. The proposed improved algorithm calculates the profit weighing factor. This factor is helpful to know the profit associated with various frequent itemsets generated.

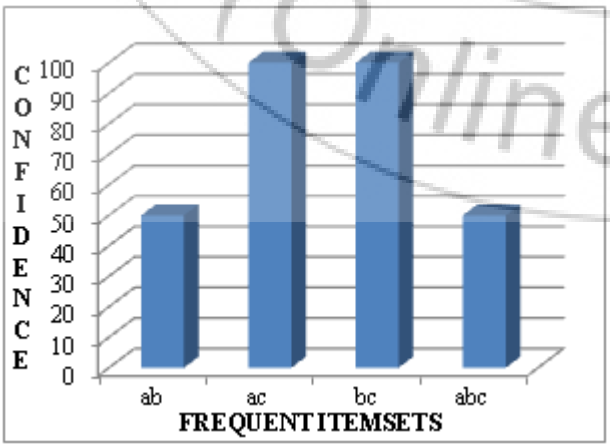

Figure 4.5: Frequent itemsets from dataset associated with confidence

The above graph shows that the frequent itemsets 'ac' and 'bc' have 100% confidence. Hence these two itemsets are highly valuable to the seller.

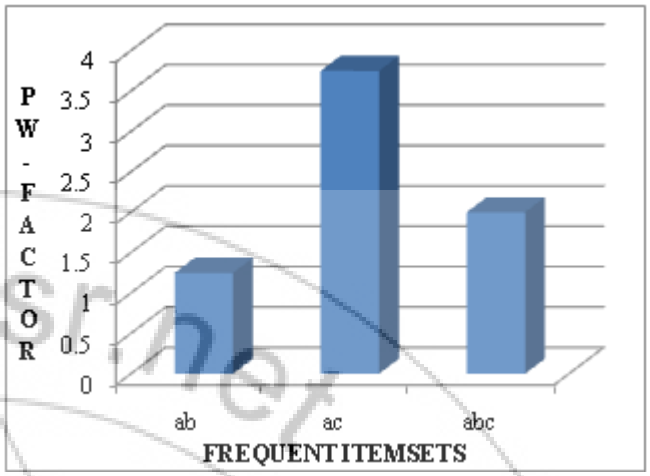

Figure 4.6: Frequent itemsets from dataset associated with P-W factor

The above graph shows that the frequent itemset 'ac' is providing the maximum profit. Though 'bc' is the frequent itemset which is providing 100% confidence with the classical Apriori algorithm but with the proposed methodology, the itemset 'bc' is providing a profit that is below the profit weighing threshold. Hence the proposed improved algorithm proves to be very valuable to the seller and thus increase the total profit to the business. The proposed algorithm is applied to many other datasets to test its validity. The results were appreciable in each of the cases and thus increasing the performance.

5. Conclusion and Future Scope

The conclusion to this work is that we can figure out the association rule by extracting out the frequent patterns from the large transactional database. Apriori algorithm is applied on the transactional database. By using measures of Apriori algorithm, frequent itemsets can be generated from the database. But the Apriori algorithm is associated with certain limitations like scanning time, memory optimization, candidate generation, etc.

Although several different strategies have been proposed to tackle efficiency issues, they are not always successful. Recently, the data mining community has turned to the mining of interesting association rules to facilitate business development by increasing the utility of an enterprise. It proposed an approach which focuses on the utility, significance, quantity and profit of individual items for the mining of novel association patterns. The mined interesting association patterns are used to offer valuable suggestions to an enterprise for intensifying its business utility.

However, in future this can be improved into a new methodology, which will output only one or two frequent itemsets that are giving maximum profit to the enterprise so that no confusion exists related to the selection of profitable itemset among large number of frequent itemsets. This proposed algorithm works for smaller number of items which

in the future can be enhanced to work for larger number of items.

References

[1] Ms. Arti Rathod, Mr. Ajaysingh Dhabariya & Mr. Chintan Thacker,(Sep. 2013) " A review on Association Rule Mining and Improved Apriori Algorithms", International Journal of Scientific Research in Computer Science , vol. 1 ,no. 11.
[2] Frawley, W., Piatetsky- Shapiro, G., Matheus, (1992). "Knowledge Discovery in Databases: An Overview", AI Magazine, fall 1992, pp. 213-228.
[3] Mamta Dhanda ,(July 2011) "An efficient approach to extract frequent patterns from transactional database" , International Journal of Engineering Science & Technology , vol. 3,no. 7 .
[4] Irena Tudor, Universitatea Petrol-Gaze din ploeiesti, (2008)"Association Rule Mining as a Data Mining Technique", Bd.Bucuresti 39, ploeiesti,Catedra de Informatica, Vol-LX,No.1.
[5] Wei Zhang ,Honghzi Liao and Na Zhao (2008), "Research On The Frequent Pattern Growth Algorithm about Association Rule Mining", International Seminar on Business and Information Management.
[6] Du Ping and Gao Yongping (2010), "A New Improvement of Apriori Algorithm For Mining Association Rules", International Conference on Computer Application and System Modeling.
[7] Dr. Dhanabhakyam, Dr.M.Punithavalli, Dr.SNS college of Arts and Science, "The Survey on Data Mining Algorithm for market basket analysis", Global Journal of Computer Science and Technology (IJCSEIT), Vol.11,Issue 11 Version 1.0, July 2012,ISSN:0975-4172.
[8] Sheila A.Abaya, "Association Rule Mining based on Apriori algorithm in minimizing candidate generation", International Journal of Scientific & engineering Research, Vo1-3, Issue-7,July 2012,ISSN 2229-5518.
[9] Raorane A.A ,Kulkarni R.V and Jitkar B.D "Association Rule- Extracting Knowledge Using Market Basket Analysis" ,Dept. of Computer Science,Vivekanand College, Tarabai Parkkolhapur, Research Journal of Recent Sciences, Vo11(2)19-27,Feb. 2012.
[10] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules", In Proceedings of 20th International Conference on Very Large Data Bases, Santiago, Chile, pp. 487–499, September 1994.
[11] Pratibha Mandave, Megha Mane, Prof. Sharada Patil (2013),"Data mining using Association rule based on APRIORI algorithm and improved approach with illustration", International Journal of Latest Trends in Engineering and Technology ,Vol. 3 Issue2, ISSN: 2278-621X.
[12] D.I.Khan (2012), "Research on Association Rule Mining", Advances in Computational Mathematics and Its Applications, Vol. 2, No. 1, 2012, ISSN 2167-6356.
[13] Abhang Swati Ashok and JoreSandeep S. (2014), "The Apriori algorithm: Data Mining Approaches is to find frequent item sets from a transaction dataset", International Journal of Innovative Research in Science, Engineering and Technology, Volume 3, Special Issue 4.
[14] Lu, S., Hu, H., Li, F, "Mining Weighted Association Rules", Intelligent Data Analysis, vol.5 , no. 3, pp.211 – 225, August 2001.
[15] María N. Moreno, Saddys Segrera and Vivian F. López, " Association Rules: Problems, solutions and new applications", Universidad de Salamanca, Plaza Merced S/N, 37008, Salamanca, 2004.
[16] Mamta Dhanda, Sonali Guglani and Gaurav Gupta (2011) , "Mining Efficient Association Rules Through Apriori Algorithm Using Attributes" , International Journal of Computer Science and Technology , Vol. 2, Issue 3.
[17] Ms Shweta and Dr. Kanwal Garg (2013) , "Mining Efficient Association Rules Through Apriori Algorithm Using Attributes and Comparative Analysis of Various Association Rule Algorithms" , International Journal of Advanced Research in Computer Science and Software Engineering , Volume 3, Issue 6.
[18] M. S. Chen, J. Han, P.S. Yu, "Data mining: an overview from a database perspective", IEEE Transactions on Knowledge and Data Engineering, vol. 8, no.6, pp. 866–883, 1996.
[19] Saeed Farzi, Ahmad Baraani Dastjerdi, " Data Quality Measurement using Data Mining", International Journal of Computer Theory and Engineering, Vol. 2, No. 1 February, 2010, ISSN: 1793-8201.
[20] Sotiris Kotsiantis and Dimitris Kanellopoulos (2006), "Association Rules Mining: A Recent Overview", International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82.
[21] Goswami D.N., Chaturvedi Anshu and Raghuvanshi C.S. (2010), "An Algorithm for Frequent Pattern Mining Based On Apriori", International Journal on Computer Science and Engineering, Vol. 02, No. 04, 942-947
[22] Jyothi Pillai (2011) , "User centric approach to itemset utility mining in Market Basket Analysis" , International Journal on Computer Science and Engineering, Vol. 3 No. 1.

Author Profile

Reeti Trikha is a student of Computer Science & Engineering in RIMT-IET, Mandi Gobindgarh affiliated to Punjab Technical University, Jalandhar, Punjab, India. She is pursuing her M. Tech research.

Er. Jasmeet Singh is an Assistant Professor in the Computer Science Department in RIMT-IET, Mandi Gobindgarh. He received his M.Tech degree from Thapar University, Patiala, Punjab, India. His research interest is in Computer Networks.