



Association Rules Mining Using Effective Algorithm: A Review

Rajdeep Kaur Aulakh

Department of Computer Science and Engineering
DAVIET, Jalandhar, Punjab, India

Abstract: Association rule mining is one of the most popular techniques of data mining methods whose aim is to extract associations among sets of items in transaction databases. However, mining association rules often results in a very large number of found rules, leaving the database analyst with the task to go through all the association rules and discover interesting ones. In this paper mining association rules has attracted a lot of attention in the research community. Several techniques for efficient discovery of association rules have appeared. However, with the increase in the size of the databases and for efficient decision making, selective marketing, market basket analysis, catalogue marketing industry etc. To reduce the limitation of Apriori algorithm of generating large number of association rules, we proposed an algorithm in this research work. In the proposed method, initially we applied Apriori algorithm in order to generate frequent item-sets and then frequent item-sets are used to generate association rules.

Keywords- Association rule mining, Apriori algorithm, FP-Growth Algorithm.

I. INTRODUCTION

Mining association rules is one of the several data mining tasks, has a big share in the data mining research. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transactional databases or other data repositories. This is attributed to its wide area of applications. Applications of association rule mining span a wide area of business from market basket analysis to analysis of promotion and catalogue design, and from designing store layout to customer segmentation based on buying patterns. Other applications include health insurance, fraudulent discovery and loss-leader analysis, telecommunication networks market and risk management, inventory control etc. Various association mining techniques and algorithms are briefly introduced and compared later. Association rule mining has the same challenges which are being faced by data mining.

1.1 IMPORTANCE OF ASSOCIATION RULE MINING

The importance of association rule mining is as follows:

1. The association rule mining helps in finding particular relationships between various data elements of the large database i.e. database having a large number of records (108- 1012 bytes).
2. Association rule mining helps in the classification of data.
3. It helps to search useful information and knowledge that can enhance the business or scientific operations.
4. To provide better and efficient methods to analyze the data. It can handle data of high dimensionality.
5. Increased competition for customers requires availability of information on demand.
6. Association rule mining helps in finding such required useful information.
7. Helps in finding the outlier entries, which may be useful in some cases such as fraud detection.

1.2 WHAT IS AN ASSOCIATION RULE:

The association rules represent the associations between the data variables. An association rule is an implication of the form written below [1].

$X \rightarrow Y$ [Support= S%, Confidence=C%], where $X, Y \subset I$ and $X \cap Y = \Phi$, and

I is an Itemset.

X is called as the Antecedent or body and

Y is called as Consequent or head of the rule.

Each rule has two measures of value support and confidence. The computation of support and confidence can be defined by the following equations:

Support ($X \rightarrow Y$) = P (XUY)

Confidence ($X \rightarrow Y$) = P(Y/X) = support_count (XUY) / support_count(X)

Where support S is the probability that rule contains {X, Y} and confidence C is the conditional probability that specify the C% of the transaction of database considered must specify $X \rightarrow Y$. Minimum support and Minimum confidence are needed to eliminate the unimportant association rules. The association rule holds iff it has the support and confidence value greater than or equal to minimum support (min_sup) and minimum confidence (min_conf) threshold value. An example of calculating support and confidence for transactional database given in Table 1.1 is described below:

Table 1.1: Example of Transactional Database

Customer	Item Purchased
(A)	Item Purchased
(B)	
1	Pizza Coke
2	Burger Sprite
3	Pizza Sprite
4	French Fries Coffee

If A is “purchased pizza” and B is “purchased soda” then

$$\text{Support} = P(A \text{ and } B) = \frac{1}{4}$$

$$\text{Confidence} = P(B / A) = \frac{1}{2}$$

Confidence does not measure if the association between A and B is random or not.

1.3 ASSOCIATION RULE MINING PROCESS

Association rule mining is a two-step process:

1. Find All Frequent Item-sets: First find all the sets of items whose support count value is equal to or more than minimum support count value. All these item sets are termed as frequent itemsets.
2. Generate strong association rules from the frequent itemsets: Second, for each frequent itemsets generate the association rules that have confidence value more than or equal to minimum confidence value.

Once the frequent itemsets from transactions in a database have been found, it is straightforward to generate the strong association rules from the frequent itemsets. This can be done by using the equation for confidence.

Based on this, association rules can be generated as follows:

For each frequent itemset L, generate all nonempty subsets of L.

For every non empty subset S of L, output the rule “S \square (L-S)”

If $(\text{support_count}(L)) / (\text{support_count}(S)) \geq \text{min_conf}$

Where min_conf is the minimum confidence threshold.

Because the second step is much less costly than the first, the overall performance of algorithm for mining association rules is determined by the first step. Different methods follow the different approach to generate the frequent itemsets. Once the frequent itemsets are found, generation of association rules from frequent patterns is easier.

1.4 VARIOUS METHODS OF ASSOCIATION RULE MINING

Various methods of association rule mining for finding the frequent itemsets are described below:

1.4.1 Apriori Algorithm

Apriori algorithm [1], [2] is one of the classical algorithms proposed by R. Srikant and R. Agrawal in 1994 for finding frequent patterns for boolean association rules. Apriori employs an iterative approach known as level-wise search, where k-itemsets are used to explore (k+1)- itemsets. First, the set of frequent 1-itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted by L1. Next, L1 is used to find L2, the set of frequent 2-itemsets, which is used to find L3, and so on, until no more frequent k-itemsets can be found. The finding of each Lk requires one full scan of the database. The algorithm is executed in two steps: First, it retrieves all the frequent itemsets from the database by considering those itemsets whose support is not smaller than the minimum support (min_sup). Secondly, it generates the association rules satisfying the minimum confidence (min_conf) from the frequent itemsets generated in first step. The first step consists of join and pruning actions. While joining, the candidate set Ck is produced by joining Lk-1 with itself and pruning of the candidate sets is done by applying the Apriori property i.e. all the non-empty subsets of a frequent itemset must also be frequent.

1.4.2 FP-Growth Algorithm

FP-growth algorithm [1], [3], [4] proposed by Jiawei Han finds the association rules more efficiently than Apriori algorithm without the generation of candidate itemsets. Apriori algorithm requires n+1 scans, where n is the length of the longest pattern. FP-growth algorithm requires only two scans of the database to find frequent patterns. FP-growth algorithm adopts divide and conquer strategy. First, it construct a FP-tree [4] using the data in transactional database and then mines all the frequent patterns From FP-tree. After mining of frequent patterns the association rules can be generated easily.

Applications of Association Rule Mining

Following are some of the applications of association rule mining:

1.4.2.1 Market Basket Analysis

A typical and widely-used example of association rule mining is market basket analysis. For example, data are collected using bar-code scanners in supermarkets. Such ‘market basket’ databases consist of a large number of transaction records. Each record lists all items bought by a customer on a single purchase transaction. Managers would be interested to know if certain groups of items are consistently purchased together. They could use this data for adjusting store

layouts (placing items optimally with respect to each other), for cross-selling, for promotions, for catalogue design and to identify customer segments based on buying patterns [6].

For example, if a supermarket database has 100,000 point-of-sale transactions, out of which 2,000 include both items A and B and 800 of these include item C, the association rule "If A and B are purchased then C is purchased on the same trip" has a support of 800 transactions (alternatively $0.8\% = 800/100,000$) and a confidence of 40% ($=800/2,000$).

1.4.2.2 Medical Diagnosis

Applying association rules in medical diagnosis can be used for assisting physicians to cure patients. The general problem of the induction of reliable diagnostic rules is hard because theoretically no induction process by itself can guarantee the correctness of induced hypotheses.

Practically diagnosis is not an easy process as it involves unreliable diagnosis tests and the presence of noise in training examples. This may result in hypotheses with unsatisfactory prediction accuracy which is too unreliable for critical medical applications.

Serban [7] has proposed a technique based on relational association rules and supervised learning methods. It helps to identify the probability of illness in a certain disease. This interface can be simply extended by adding new symptoms types for the given disease, and by defining new relations between these symptoms.

II. RELATED WORK

Many surveys have been conducted on previously developed optimization techniques. Logical organization of this literature survey proved to be a vital task for filling the gap between researches recently to improve performance of the association rule mining.

Shanta Rangaswamy and G.Shobha [5] presented a method in which genetic algorithm [6] is applied over the rules fetched from Apriori association rule mining. By using Genetic Algorithm the proposed system can predict the rules which contain negative attributes in the generated rules along with more than one attribute in consequent part. The goal of generated system was to implement association rule mining of data using genetic algorithm to improve the performance of accessing information from databases (Log file) maintained at server machine and to improve the performance by minimizing the time required for scanning huge databases maintained at server machines.

Sanat Jain, Swati Kabra[6] presented an Apriori-based algorithm that is able to find all valid positive and negative association rules in a support confidence framework. The algorithm can find all valid association rules quickly and overcome some limitations of the previous mining methods. Authors have designed pruning strategies for reducing the search space and improving the usability of mining rules, and have used the correlation coefficient to judge which form association rule should be mined.

Jun Gao [7] presented a new method called as MFP algorithm is an improvement over FP-growth algorithm. FP-growth algorithm requires two database scans one for construction of table L and second for construction of FP-tree. But in case of MFP algorithm only one database scan is required. MFP Algorithm consist of two main steps: first, construction of MFP-tree and mining of frequent patterns from MFP-tree. In MFP-tree each node expect the root node and leaf node has two enteries. Support count value of node and pointer to the next node in MFP-tree. The results of this paper has shown that MFP algorithm requires less time and can find the frequent patterns by scanning the database only once. This algorithm can be applied to any situation where FP-growth or Apriori algorithm is suitable.

Li Juan and Ming De-ting [8] proposed a new method called QFP algorithm. It is an improvement over FP-growth algorithm. QFP algorithm requires only one database scan to convert the transaction database into QFP-tree after data preprocessing. Then directly generates the association rules from the QFP-tree without looking the transaction database. This algorithm works in two steps: First, construction of QFP-tree and then, mine the QFP- tree to obtain the frequent patterns. The experimental result of QFP algorithm has shown that time efficiency of the QFP algorithm is higher that that of FP-growth algorithm. The QFP algorithm can be applied to any situation which is suitable for FP-growth or Apriori algorithm as the input to QFP is same as that of FP-growth or Apriori algorithm.

Zhi Liu, Mingyu Lu, Weiguo Yi and Hao Xu [9] presented a new method for association rule mining algorithm based on coding and constraint uses the properties of Apriori algorithm and makes some improvement based on it. The algorithm uses the sub-block coding method for properties and applies constraints for antecedent and consequent of the rules. In this method the attribute value is divided into decision attributes and non-decision attributes. Decision attribute appears in the antecedent of the association rule and non-decision attributes can only appear in the consequent of the rule. The results has shown that the new method reduces the number of candidate itemset generated and also reduces the number of times the database is scanned.

Wanjun Yu, Xiao Chun Wang, Fangyi Wang, Erkang Wang and Bowen Chen [10] proposed a novel algorithm called as Reduced Apriori Algorithm with Tag (RAAT). The proposed algorithm reduces the number of candidate itemset produced in pruning operation of C2 and thus improves the efficiency and saves time. The algorithm RAAT optimize subset operation by using transaction tag to speed up support calculation. The experimental results of this paper shows that the RAAT algorithm gives better result in terms of candidate generation and counting the support using database as compare classical Apriori algorithm.

Dongme Sun, Shaohua Teng, Wei Zhang and Haibin Zhu [11] presented a new algorithm to improve the effectiveness of Apriori algorithm. In this algorithm the researcher used the combination of reverse and forward scan of database to find the maximal frequent itemset [1]. In this algorithm they used the concept of dynamic itemset counting and use the barrel structure [2] to store all the frequent itemsets. In this first, Lk maximal frequent itemset is found along with its support.

After this next frequent itemsets are mined i.e. L_{k-1} and their respective support value is counted by using database D. Similarly all the frequent itemset are mined in this way. Then all these frequent itemsets are placed in bit -matrix [12] to count their respective support values. The results of this shows that the improved barrel structure method requires very less time for scanning the database as compare to Apriori algorithm and saves the space as it does not produce large number of candidate itemsets.

III. APRIORI ALGORITHM

Apriori algorithm [1], [2] is one of the classical algorithms proposed by R. Srikant and R. Agrawal in 1994 for finding frequent patterns for boolean association rules. Apriori employs an iterative approach known as level-wise search, where k-itemsets are used to explore (k+1)- itemsets. First, the set of frequent 1-itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted by L₁. Next, L₁ is used to find L₂, the set of frequent 2-itemsets, which is used to find L₃, and so on, until no more frequent k-itemsets can be found. The finding of each L_k requires one full scan of the database.

The algorithm is executed in two steps:

1. Prune and Join: First, it retrieves all the frequent itemsets from the database by considering those itemsets whose support is not smaller than the minimum support (min_sup). The first step consists of join and pruning actions. While joining, the candidate set C_k is produced by joining L_{k-1} with itself and pruning of the candidate sets is done by applying the Apriori property i.e. all the non-empty subsets of a frequent itemset must also be frequent.

Algorithm: The basic algorithm of mining association is given as follows:

Let I = {I₁, I₂...I_n} be a set of item and

D = {T₁, T₂ ...T_n} be a set of transaction

Where t_i is a set of transaction t_i ∈ I, An association rule is transaction of the form X → Y

Where X, Y ⊂ I and X ∩ Y = ∅. The rule X → Y holds in the set D with Support and Confidence.

An example of All Electronics Transactional Database D [1] is presented below in Table 3.1 to specify the process of Apriori algorithm. Let min_sup=2 and min_conf as 70%.The process of generating frequent itemsets by Apriori algorithm is shown below in Table 1.1

Table 1.1: All Electronics Transactional Database (D)

TID	Itemsets
T001	I1, I2, I5
T002	I2, I4
T003	I2, I3
T004	I1, I2, I4
T005	I1, I3
T006	I2, I3
T007	I1, I3
T008	I1, I2, I3, I5
T009	I1, I2, I3

The pseudo code for generation of frequent itemsets is given below in Figure 1.2.

```

Ck: The set of candidate itemsets of size k
Lk: The set of frequent itemsets of size k
{
L1= frequent 1-itemsets
For (k=2; Lk-1! =NULL; k++)
{
Ck=Join Lk-1 with Lk-1 to generate Ck;

Lk= Candidate in Ck with support greater than or equal to
minimum support;
L=L U Lk // L is a set containing all frequent itemsets
}
End;
Return L;
}
    
```

Figure 1.2: Pseudo code of Apriori Algorithm

IV. PROPOSED ALGORITHM

Applied Algorithm

Algorithmic Structure: The proposed method for generating association rule by GA is as follows:

Step 1: Start

Step 2: Load a sample of records from the database that fits in the memory.

Step 3: Apply Apriori algorithm to find the frequent item sets with the minimum support. Suppose A is set of the frequent item set generated by Apriori algorithm.

Step 4: Set $Z = \emptyset$ where Z is the output set, which contains the association rule.

Step 5: Input the termination condition of GA.

Step 6: Represent each frequent item set of A as a binary string using the combination of representation.

Step 7: Select the two members from the frequent item set using Roulette Wheel sampling method.

Step 8: Apply the crossover and mutation on the selected members to generate the association rules.

Step 9: Find the fitness function for each rule $X \square Y$ and check the following condition.

Step 10: If (fitness function > min confidence)

Step 11: Set $Z = Z \cup \{X \square Y\}$

Step 12: If the desired number of generations is not completed, then go to Step 3.

Step 13: Stop.

V. CONCLUSION

Data mining and knowledge discovery are new emerging disciplines with important applications in Science, engineering, health care, education and business. Association rule mining is one of the key fields in data mining. Many researchers are trying to develop efficient methods to find the frequent patterns and to optimize association rules. Association rule mining is an important topic in data mining and receiving increasing attention. An efficient algorithm for optimization of association rule mining has been proposed in this research work. Various association rule mining algorithms such as Apriori suffers from limitation of large number of association rules generation. An efficient method is developed in this research work to find the minimized number of association rules.

REFERENCES

- [1] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan- Kaufmann Publishers, 2000.
- [2] R. Agrawal, R. Srikant, "Fast Algorithm for Mining Association Rules", Proc. of the Int. Conf on Very Large Database, pp. 487- 499, 1994.
- [3] J. Han, J. Pei, Y. Yin, "Mining Frequent Patterns without Candidate Generation". Proc. of the ACM SIGMOD Int. Conf. on Management of Data, pp.1-12, 2000.
- [4] J. Han, J. Pei, Y. Yin and R. Mao, "Mining Frequent Patterns without Candidate Generation: A Frequent Pattern Tree Approach", In Data mining and Knowledge Discovery, Vol. 8, pp.53-87, 2004.
- [5] S. Rangaswamy, Shobha G., "Optimized Association Rule Mining Using Genetic Algorithm," Journal of Computer Science Engineering and information Technology Research (JCSEITR), Vol.2, Issue 1, pp 1-9, 2012.
- [6] S. Jain, S. Kabra. "Mining & Optimization of Association Rules Using Effective Algorithm," International journal of Emerging Technology and Advanced Engineering (IJETA), Vol.2, Issue 4, 2012.
- [7] Jun Gao, "A New Association Rule Mining algorithm and Its Applications", IEEE 3rdInt. Conf. on Advanced Computer Theory and Engineering (ICACTE), vol 5, pp. 122-125,2010.
- [8] Li Juan and Ming De-ting, "Research of an association rule mining algorithm based on FP tree", IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS), Vol. 1, pp. 559-563,2010.
- [9] Zhi Liu, Mingyu Lu, Weiguo Yi, and Hao Xu, "An Efficient Association Rules Mining Algorithm Based on Coding and Constraints", Proceedings of the 2nd International Conference on Biomedical Engineering and Informatics, pp. 1-5, 2009.
- [10] Wanjun Yu , XiaochunWang, and Fangyi Wang, "The Research of Improved Apriori Algorithm for Mining Association Rules", 11th IEEE International Conference on Communication Technology Proceedings, pp. 513-516, 2008.
- [11] Dongme Sun, Shaohua Teng, Wei Zhang and Haibin Zhu, "An Algorithm to Improve the Effectiveness of Apriori Algorithm", Proc. of 6th IEEE Int. Conf. on Cognitive Informatics", pp. 385-390, 2007.
- [12] Chin-Feng Lee and Tsung-Hsien Shen, "An FP-Split Method for Fast Association Rule Mining", Proc. of IEEE 3rd International Conference on Information Technology: Research and Education, June 27-30, pp. 459-463, 2005.
- [13] Deepa, M. Kalimuthu. "An Optimization of Association Rule Mining Algorithm using Weighted Quantum behaved PSO", International Journal of Power Control Signal and Computation (IJPCSC), Vol.3, 2012.
- [14] S. Dehuri, R. Mall, "Mining Predictive and Comprehensible Rules Using A Multi-Objective Genetic Algorithm", Advance Computing and Communication (ADCOM), India, 2004.
- [15] J. Arunadevi, V. Rajamani. "Optimization of Spatial Association Rule Mining using Hybrid Evolutionary algorithm." International Journal of Computer Applications Vol. 1, Issue 19, 2010.
- [16] Y. Cheung, A. Fu, "Mining Frequent Item sets without Support Threshold: With and Without Item Constraints", IEEE Transactions on Knowledge and Data Engineering, Vol. 16, pp. 1052-1069, 1999.