"An Efficient Analysis of Execution Time and Association Rules for Apriori Algorithm with Greedy Method for different values of Minimum Support"

Ms. Arpita Lodha[1], Vishal Shrivastava[2]
[1] M. Tech Scholar, Dept. of CS/IT, ACEIT, Arya Group of Colleges, Jaipur, Rajasthan, India
[2] Associate Professor, Dept. of CS/IT, ACEIT, Arya Group of Colleges, Jaipur, Rajasthan, India

ABSTRACT

The problem of deriving associations from database communities has gathered huge attentions. Mining Association rules is one of the main research area of data mining techniques which is used in many real life applications. There are several mining algorithms of deriving association rules. One of the most popular is Apriori based on this algorithm, this paper mention the limitations of the classical Apriori Algorithm and presents an analysis of execution time and number of association rules generated with greedy method for different values of minimum support keeping minimum confidence constant.

Keywords: *Apriori, Frequent itemsets, Minimum Support, Confidence, Greedy Algorithm.*

1 INTRODUCTION

Data mining is a method that combines traditional data analysis methods with sophisticated architecture, algorithms, and techniques for processing huge volume of structured, semi-structured and unstructured data set. It has also opened the door for exploring and analysing new types of data and for analysing old type of data in newer ways with exciting opportunities. Data mining draws ideas from statistics, machine learning, pattern recognition system and data base system.

The process which finds the interesting associations among large set of data items is called association rule mining[1].An association rule identifies attribute value conditions which are frequently occur in given transactional database. Market Basket analysis is an application of association rule mining.

Association analysis is the discovery of what are commonly called association rules. It studies the frequency of items occurring together in transactional databases, and based on a threshold called minimum support, identifies the frequent item sets. Another threshold confidence, which is the conditional probability that an item appears in a transaction when another item appears, is used to pinpoint association rules.

2 LITERATURE SURVEY

In this section author has discussed some research papers which had been previously undertaken in the field of association rule mining.

Guofeng Wang, Xiu Yu, Dongbiao Peng, Yinhu Cui, Qiming Li, [2] In this Paper, Apriori Algorithm for association rules mining, and make some improvement for this algorithm based on the features of cutting database. Apriori Algorithm is improved to mine association rules in cutting database. The results show that the Apriori algorithm can be efficiently used in cutting data mining, and improved algorithm can achieve expected effect better than traditional algorithm. Cutting data mining is an important method to increase efficiency, discover hidden knowledge in cutting database, and provide guidance for cutting decisions.

Wang Peiji, Shi Lin, Bai Jinniu, Zhao Yulin, [3] In this Paper, Mining association rules is an important topic. Apriori Algorithm submitted by Agrawal and R. Srikant in 1994 is the most effective Algorithm. Aimed at two problems of discovering frequent itemsets in a large database and mining association rules from frequent

itemsets, I make some research on mining frequent itemsets algorithm based on Apriori Algorithm and mining association rules algorithm based on improved measure system. Mining association rules algorithm based on sup-port, confidence and interestingness is improved, aiming at creating interestingness useless rules and losing useful rules. Useless rules are cancelled, creating more reasonable association rules including negative items.

D.S. Rajput, R.S. Thakur, G.S.Thakur,[4] In this Paper, he has used concept of fuzzy approach, in uncertain data set and used multiple minimum supports to find fuzzy association rules in a given uncertain data set. This works well with problem of uncertain data relation-ships, which are represented by fuzzy set concepts. The proposed approach can thus generate large frequent item sets and then derive fuzzy association rules from uncertain data set. The advantage of fuzzy association rule mining is that, the algorithm performs scanning our data set only one time. By this also we have to do less effort for the calculation of various relations. This approach scale well when handling large amount of dense data. As ongoing work, we are investigating ways to further reduce the data structure problems, and generated frequent patterns will be useful for clustering.

R. Chang and Z.liu, [5] In this Paper, he has used hash structure which optimizes 2-items generation which improves the save time and space. By using hash table instead of hash tree the searching cost is reduced. The 2-member candidate itemsets are directly generated using a hash function.

3 CLASSICAL APRIORI ALGORITHM

Some common terminologies are used in Apriori Algorithm [6].

- **Minimum support**-An itemset containing i items is called i-itemset. The percentage of transactions that contain an itemset is called the itemset's support.
- **Confidence**-Confidence is the ratio of the number of the transactions that include all items in the consequent and antecedent to the total number of transactions that include all items in the antecedent.
- **Frequent Sets**-The itemsets which stratifies the minimum support criteria are known as frequent itemsets. It is denoted by L_i where i indicate the i-itemset.

Apriori is a Algorithm proposed by R.Agarwal and R.Shrikant in 1994[7]. Apriori employs an iterative approach known as a 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 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. To improve the efficiency of the level-wise generation of frequent itemsets, an important property called the Apriori property, presented is used to reduce the search space. Apriori property: All nonempty subsets of a frequent itemset must also be frequent. A two-step process is used to find the frequent itemsets: join and prune actions.

(a) The join step: To find L_k a set of candidate k-itemsets is generated by joining L_{k-1} with itself. This set of candidates is denoted C_k.

(b) The prune step: The members of C_k may or may not be frequent, but all of the frequent k-itemsets are included in C_k. A scan of the database to determine the count of each candidate in C_k would result in the determination of L_k (i.e., all candidates having a count no less than the minimum support count are frequent by definition, and therefore belong to L_k). To reduce the size of C_k, the Apriori property is used as follows. Any (K-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset. Hence, if any (K-1)-subset of a candidate k-itemset is not in L_{k-1}, then the candidate cannot be frequent either and so can be removed from Ck.

Limitations of Apriori

The first limitation of this algorithm is the generation of a large number of candidate itemsets. The second problem is the number of database passes which is equal to the max length of frequent itemsets

IV MODIFIED APRIORI

According to proposed work main focus is to analyse the variation in execution time and number of association rules generated in greedy Apriori Algorithm.

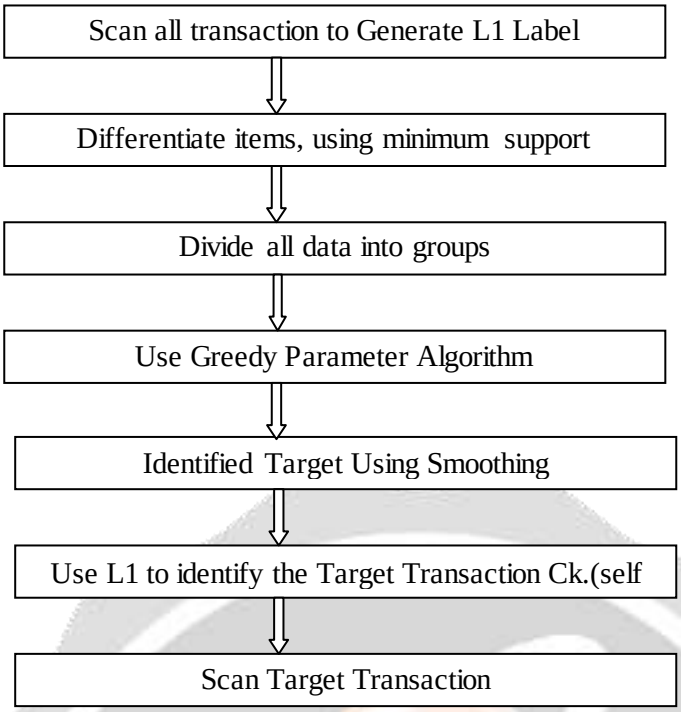

Figure: 4.1 Process of Proposed work

4.1 GREEDY ALGORITHM

Greedy is a strategy that works well on optimization problems with the following characteristics:

1. Greedy-choice property: - A global optimum can be arrived at by selecting a local optimum.
2. Optimal substructure: - An optimal solution to the problem contains an optimal solution to sub problems.

4.2 ORTHOGONAL GREEDY ALGORITHM

Orthogonal matching pursuit (OMP) algorithm has received much attention in recent years. OMP algorithm is an iterative greedy algorithm that selects at each step the column [8]. Orthogonal matching pursuit (OMP) constructs an approximation by going through an iteration process. At each iteration the locally optimum solution is calculated. This is done by finding the column vector in A which most closely resembles a residual vector r. The residual vector starts out being equal to the vector that is required to be approximated i.e. r = b and is adjusted at each iteration to take into account the vector previously chosen. It is the hope that this sequence of locally optimum solutions will lead to the global optimum solution. As usual this is not the case in general although there are conditions under which the result will be the optimum solution. OMP is based on a variation of an earlier algorithm called Matching Pursuit (MP). MP simply removes the selected column vector from the residual vector at each iteration.

$$r_t = r_{t-1} - < aOP , r_{t-1} > r_{t-1}$$

Where aOP is the column vector in A which most closely resembles r_{r-1}.OMP uses a least-squares step at each iteration to update the residual vector in order to improve the approximation. The OMP is a stepwise forward selection algorithm and is easy to implement.

5 EXPERIMENTAL RESULTS

The proposed modified algorithm was implemented using MATLAB. Moreover, the benchmark module was also used to measure the execution time of the code and number of association rules generated. There were 5 different values of minimum support and constant confidence were taken for analysis purpose and result are shown in table 5.1

Sr. No.	Minimum Support (M) and Confidence(C)	Association Rules Generated	Final Association Rules Generated	Execution Time (in Sec)
1	M=0.22; C=0.6;	50	16	10.136450
2	M=.44; C=0.6;	14	7	3.824037
3	M=.66; C=0.6;	12	6	2.961933
4	M=.88; C=0.6;	0	0	0.606136
5	M=1.1; C=0.6;	0	0	0.569182

Table 5.1 Minimum support v/s Association Rules and Execution time

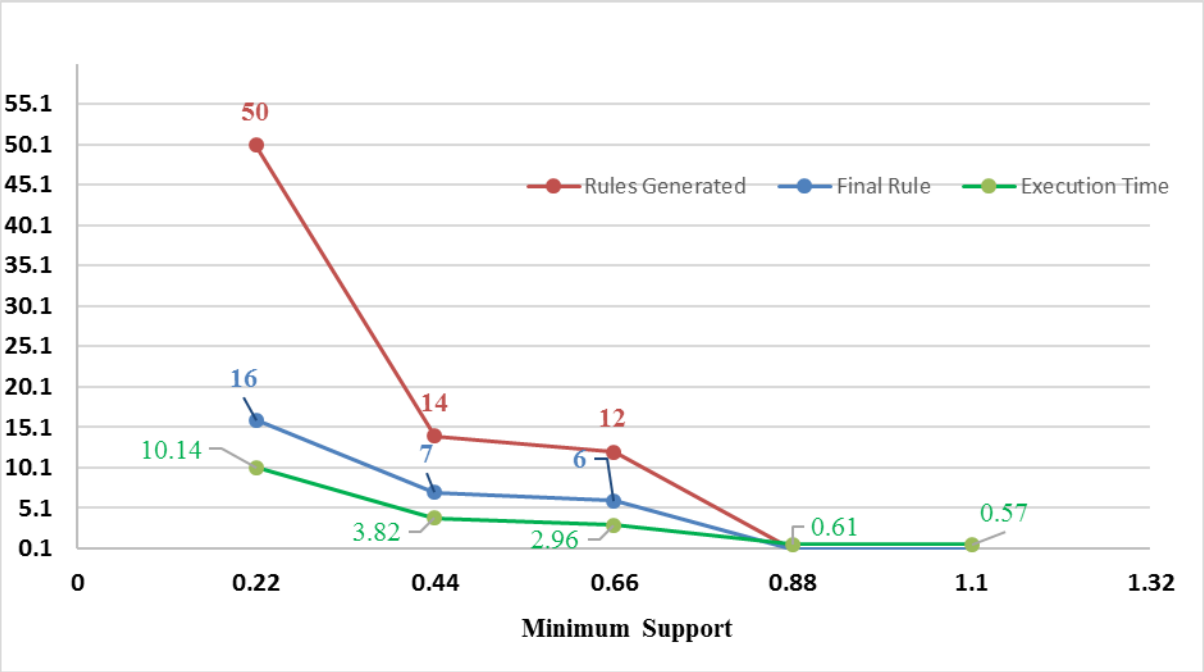

Figure 5.1 Minimum support v/s Association Rules and Execution time

VI CONCLUSION

In this paper, the execution time, number of rules and final rules is reduced by increasing the value of minimum support i.e. minimum support is inversely proportional to execution time and number of association rules.

REFERNCES

[1] Jingyao Hu, "The Analysis on Apriori Based on Interest Measure", IEEE-2012, ICCECT,978-0-7695-4881.
[2] Guofeng Wang, Xiu Yu, Dongbiao Peng, Yinhu Cui, Qiming Li, "Research of Data Mining Based on Apriori algorithm in Cutting Database", IEEE – 2010
[3] WANG Pei-ji, SHI Lin, BAI Jin-niu, ZHAO Yu-lin, "Mining Association Rules Based on Apriori Algorithm and Application", IEEE - 2009, pp.141-143.
[4] D.S. Rajput, R.S. Thakur, G.S. Thakur, "Fuzzy Association Rule Mining based Frequent Pattern Extraction from Uncertain Data", World Congress on Information and Communication Technologies, IEEE-2012,pp.709-714.
[5] R. Chang and Z. liu, "An Improved Apriori Algorithm", no. Iceoe, pp.476-478,2011.
[6] Jiawei Han, Micheline Kamber, "Data Mining, Concepts and Techniques", ISBN 978-81-312-0535-8, Elsevier India Private Limited,2006.
[7] R. Agarwal and R. Shrikant , "Fast Algorithms for mining association rules in large database. In Research report RJ 9839, IBM Almaden Research Center, San Jose, CA, June 1994.
[8] T. Tony Cai and Lie Wang, "Orthogonal Matching Pursuit for Sparse Signal Recovery With Noise", IEEE Transactions on Information Theory, Vol. 57, No. 7, July 2011.