An Efficient Enhancement of Mining Top-K Association Rule

Harshit Srivastava*
M.Tech Scholar
Suresh Gyan Vihar University
Jaipur, India

Virendra Kumar
Assistant Professor
Suresh Gyan Vihar University
Jaipur, India

Savita Shiwani
Associate Professor
Suresh Gyan Vihar University
Jaipur, India

Abstract: Data mining on huge databases has been a major issue in research area, due to the problem of analyzing large volumes of data using traditional OLAP tools only. This type of process implies much computational power, disk I/O and memory, which can be used only by parallel computers. So, depending on the selection of the parameters (the minimum support and minimum confidence), current algorithms can be slow and produce an extremely large number of results or produce very less results, omitting useful information. This is really a major problem because in practice users don’t have much resource for analyzing the results and have to discover a certain amount of results in a limited time. To address this problem, we propose a unique technique to mine top ranked data from a data set. The algorithm uses a new method for generating association rules. The algorithm has unique and best performance and feature of scalability, which is a beneficial alternative to classical Association rule mining algorithms when the user wants to control the number of rules generated.

Keyword: Association Rule Mining, Data mining, Association rule learning, Top-k rules, Confidence

I. INTRODUCTION

Data mining includes the use of data analysis tools to find out hidden data, un-known, valid patterns and relationships in large databases. These tools contain mathematical algorithms, statistical models, and machine learning methods like decision trees or neural networks. Resultantly, data mining comprise of more than managing and collecting data, it also consists of analysis and prediction. Data mining can be used on data represented in textual, multimedia or quantitative forms. Applications of Data mining can use different-2 parameters to check the data. Data mining applications consist of sequence or path analysis ( such type of patterns where events leads to each other, such as the birth of a child and purchasing diapers), association (such type of pattern where events are correlated to each other, such as purchasing a bread and butter), classification (recognition of new patterns, such as coincidences between plastic sheeting purchases and duct tape purchases ), clustering (grouping the previously un-known data in to class of similar data object, such as students and their sections).

Association rule learning in data mining is a good method for finding interesting relations between huge data sets. It is used to recognize strong rules founded in databases using distinct measures of interestingness. For example, the rule \{butter, jam\} $\Rightarrow$ \{bread\} would indicate that if a customer buys butter and jam together, he or she also buy bread. Such type of information can be used for decisions making in marketing like product marketing and the price promotion.

By considering above it is cleared that association rules are applied today in many areas like bioinformatics, Web usage mining and intrusion detection. Association rule learning does not consider the order of items either within a transaction or across transaction.

II. RELATED WORK

Apriori [4,6,8] its variants are a great improvement in the history of association rule mining. Apriori algorithm was first proposed by Agrawal in [Agrawal and Srikant 1994]. The AIS is just a straight forward approach that requires many passes over the database, generating many candidate item sets and storing counters of each candidate . In later stages, most of the candidates found in step 1, were infrequent. Apriori is better in the process of candidate generation for two reasons; Apriori employs a different candidate’s generation method and a new pruning technique. The Apriori algorithm has the drawback of scanning the whole data bases many times. After modification in apriori algorithm some new algorithm has been designed. Generally there were two approaches: one is to reduce the number of passes over the whole database or replacing the whole database with only part of it based on the current frequent item sets, another approach is to explore different kinds of pruning techniques to make the number of candidate item sets much smaller. Apriori-TID and Apriori-Hybrid [Agrawal and Srikant 1994], DHP [Park et al. 1995], SON [Savesere et al. 1995] are modifications of the Apriori algorithm [9,10,11].

Most of the algorithms introduced above are based on the Apriori algorithm and try to improve the efficiency by making some modification like reducing number of scan over the database, reducing the size of the database to be...
scanned in every pass; pruning the candidates by different techniques and using sampling technique. There are two disadvantage of the Apriori algorithm. First, the process of candidate generation in apriori algorithm is very complex. It uses most of the time & space. Second, multiple passes over the database.

To break the two bottlenecks of Apriori series algorithms, some works of association rule mining using tree structure have been designed. FP-Tree [7], frequent pattern mining, is new research development in association rule mining, which overcomes the two limitations of the Apriori. In this, frequent pattern are generated without candidate generating process with two times scan over the database. FP-Tree was introduced by Han et al. By avoiding the candidate generation process and less passes over the database, FP-Tree is an order of magnitude faster than the Apriori algorithm. The FP generation process consist of two sub processes: constructing the FP-Tree, and production of frequent pattern from FP-Tree.

This algorithm has three major advantages. These are discussed as follows. First In FP-Tree database has reduced structure compared to the original database because only frequent items are considered and other useless items are omitted. Also by ordering the items according to their supports the overlapping paths appear only once with different support count. Secondly this algorithm only scans the database twice. The frequent patterns are generated by the FP-growth procedure, constructing the conditional FP-Tree which contain patterns with specified suffix patterns, frequent patterns can be easily generated as shown in above the ex- ample. Also the computation cost decreased dramatically. thirdly, this algorithm uses a divide and conquer technique that reduced the size considerably.

Some algorithms for finding top k rules are proposed in [2,5]. But they are inefficient. Every algorithm has its limitations, like FP-Tree algorithm is difficult to be used in an interactive mining system. During the interactive mining process, users may change the threshold of support according to the rules. However in FP-Tree if we change the support then whole mining process will repeat again and again. Another more limitation of the FP-Tree is that it does not suit for incremental mining.

III. PROPOSED WORK

The Top Rank algorithm takes as input a transaction database, a number n of rules that the user wants to discover, and the minconf threshold. The algorithm main idea is the following. TopRank first sets an internal minsup variable to 0. Then, the algorithm starts searching for rules. As soon as a rule is found, it is added to a list of rules LIST ordered by the support.

The list is used to maintain the top n rules found until now. Once n valid rules are found, the internal minsup variable is raised to the support of the rule with the lowest support in LIST. Raising the minsup value is used to prune the search space when searching for more rules.

Thereafter, each time a valid rule is found, the rule is inserted in LIST, the rules in LIST not respecting minsup anymore are removed from LIST, and minsup is raised to the value of the least interesting rule in LIST.

The algorithm continues searching for more rules until no rule are found, which means that it has found the top rank rules.

A) Procedure of Finding Top Ranked Data Set:
Finding Top ranked data set values from standard data set is give below:

Step 1: Start

Step 2: Read the following as input

2.1 Transaction Database
2.2 Parameter n
2.3 Minimum confidence

Step 3: Initialize minsup (minimum support) to 0

Step 4: Repeat step 5 to 6 While (There are rules)

Step 5: If supp (rule) => minsup and conf (rule) =>minconf AND then it is added to a list of rules (RULELIST).

RULELIST is a list of top n rules ordered by support.

Step 6: Set minsup = support of rule with the lowest support in RULELIST

Step 7: Stop
B) Advantage of Proposed algorithm

1) **Redundancy:**
   our new approach reduces the redundant rules so we find accurate result.

2) **Time Efficient:**
   An approach to add new mine rule to LIST give more control on program. It can be suitable for user who wants to satisfy on first result.

3) **Reduce Memory Usage:**
   As with new approach with reduce the redundant data and Mine only give rule less memory will we capture by approach as previous methods can display output when all rules will mine.

4) **Sorting of Data:**
   When output store in LIST it will be arrange in descending order by SUP value of new mine rule thus not required user to sort data again.

5) **Reduce computation:**
   It also save computation time because computation is needed to be done for only top rules.

IV. RESULT & CONCLUSION

When the data set is too large, the general association rule mining algorithm can generate an extremely large amount of rules. It takes a lot of execution time and also consumes huge memory.

The following chart is showing the execution time of different algorithm with the comparison of top-k association algorithm.

![Comparison of Execution Time with Different Data Row Value in Data Set](image)

In another case the association rule mining algorithm may generate rules with redundant data set row. In this particular case loss valuable information an user can’t choose how many rules he wants to display. To overcome these above mentioned problems, pro- posed a novel algorithm for mining top ranked data from any standard data set. Testing pro- posed algorithm on a standard data set. This data set is available under general public license (GNU).

**ACKNOWLEDGMENT**

The author express his pleasure and privilege to acknowledge and express his deep sense of gratitude to his teacher and guide, Mr. Virendra Kumar Department of Information Technology, Suresh Gyan Vihar University, Jaipur, who inspired and helped him to initiate and prepare this paper despite his busy academic schedule. The author also want to give his thanks to Mrs. Savita Shiwani for giving her valuable time and thought in giving necessary guidance.
REFERENCES

[1] Tan Kumar, “data Mining techniques”


