COMPARISON AND IMPROVEMENT OF ASSOCIATION RULE MINING WITH INTEGRATION OF A-PRIORI & FP-GROWTH ALGORITHM IN MATLAB: A REVIEW

Ankur Soni[1], Dushyant Singh[2]
[1]M.Tech –CSE, [2]Asst. Prof. at Dept of CSE, Vivekananda Global University, Jaipur

Abstract: Association rules are one of the major techniques of data mining. Association rule mining finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories. The volume of data is increasing dramatically as the data generated by day-to-day activities. Therefore, mining association rules from massive amount of data in the database is interested for many industries which can help in many business decision making processes, such as cross-marketing, Basket data analysis, and promotion assortment. The techniques for discovering association rules from the data have traditionally focused on identifying relationships between items telling some aspect of human behavior, usually buying behavior for determining items that customers buy together. All rules of this type describe a particular local pattern.

Key Word: A-Priori, FP-Growth, Association Rule Mining, MATLAB

I. INTRODUCTION

This is an important part of KDD. Data extraction usually consists of four classes; Classification, Grouping, Consideration and Learning of the Association of Associations. Data mining refers to discover a large amount of data. This is a scientific management that is relevant With the analysis of speculative data sets with the objectives of unexpected findings Create a summary of the data in novel ways for relationships and bosses the appendix includes views of data mining as an area of study and use Instead of managing the network, many domains are four basic themes, which are Data mining assistance includes:

Statistics: This can provide tools to measure the importance of the given data, Opportunities and many other tasks (for example, linear regression).

Learning machine: provides the algorithm to generate information related to the data (EGSM).

Data management and database: From the processing of massive data, an effective way to access and maintain data is necessary.

Artificial Intelligence: This helps to codify or join work Search techniques (EG Natural Network)

1.1 A-priori Algorithm

The first algorithm for mining all frequent item-sets and strong association rules was the AIS algorithm by [3]. Shortly after that, the algorithm was improved and renamed A-priori. A-priori algorithm is, the most classical and important algorithm for mining frequent item-sets. A-priori is used to find all frequent item-sets in a given database DB. The key idea of A-priori algorithm is to make multiple passes over the database. It employs an iterative approach known as a breadth-first search (level-wise search) through the search space, where k-item-sets are used to explore (k+1)-item-sets. The working of A-priori algorithm is fairly depends upon the A-priori property which states that" All nonempty subsets of frequent item-sets must be frequent" [2].

1.2 FP-Growth Algorithm

FP-tree algorithm [5, 6] is based upon the recursively divide and conquers strategy; first the set of frequent 1-itemset and their counts is discovered. With start from each frequent pattern, construct the conditional pattern base, then its conditional FP-tree is constructed (which is a prefix tree.). Until the resulting FP-tree is empty, or contains only one single path. (Single path will generate all the combinations of its sub-paths, each of which is a frequent pattern). The items in each transaction are processed in L order. (i.e. items in the set were sorted based on their frequencies in the descending order to form a list). The detail step is as follows: [6]

FP-Growth Method:

- Construction of FP-tree create root of the tree as a "null".
- After scanning the database D for finding the 1-itemset then process the each transaction in decreasing order of their frequency.
- A new branch is created for each transaction with the corresponding support.
- If same node is encountered in another transaction, just increment the support count by 1 of the common node.
- Each item points to the occurrence in the tree using the chain of node-link by maintaining the header table. After above process mining of the FP-tree will be done by Creating Conditional (sub) pattern bases:
- Start from node constructs its conditional pattern base.
- Then, Construct its conditional FP-tree & perform mining on such a tree.
- Join the suffix patterns with a frequent pattern

generated from a conditional GP-tree for achieving FP-growth.

- The union of all frequent patterns found by above step gives the required frequent item-set. In this way frequent patterns are mined from the database using FP-tree.

Mining Frequent Item-sets for Association Rules

Database has been used in business management, government administration, scientific and engineering data management and many other important applications. The newly extracted information or knowledge may be applied to information management, query processing, process control, decision making and many other useful applications. With the explosive growth of data, mining information and knowledge from large databases has become one of the major challenges for data management and mining community.

II. LITERATURE REVIEWS

[1] María N. Moreno, Saddys Segrera, Vivian F. López and M. José Polo studied on improving the quality of association rules by preprocessing numerical data. They stated that Many data mining problems require dealing with continuous numerical attributes, which must be split into intervals of values in order to obtain comprehensible results. This process, named attribute binning or discretization, is carried out either in preprocessing, or embedded within the induction algorithm, depending on the method. Some learning methods, such as decision tree, enclose the partitioning procedure; however, association rule algorithms do not usually include it. The paper deal with the problem of finding useful association rules from software project management data. The main drawbacks in this application field are the treatment of continuous attributes and the difficulty to obtain domain knowledge in order to evaluate the interestingness of the association rules. [2] Trupti A. Kumbhare (Research Student, DYPIET) and Prof. Santosh V.Chobe (Associate Professor, DYPIET, Pimpri, Pune, India) gave their views on an overview of association rule mining algorithms in 2014. They stated that the Data is important property for everyone. Large amount of data is available in the world. There are various repositories to store the data into data warehouses, databases, information repository etc. This large amount of data needs to process so that we can get useful information. [3] Zainab Darwish, Mousa Al-Akhras and Mohamed Habib studied on use filtering techniques to improve the accuracy of association rules in 2017. They stated that Association rules' learning is a machine learning technique used to find interesting relations among data items and is a base to build an association rules classifier. The accuracy of the classifier highly depends on the quality and accuracy of data items. This accuracy can be affected negatively by noisy instances and this may lead to classification overfitting. This work investigates overcoming this problem by applying DROP3 or all KNN filtering algorithms to the datasets prior to generating association rules and building a classifier. Several experiments and comparisons were conducted to test the accuracy of the above filtering algorithms. [4] Xiao-Feng Gu, Xiao-Juan Hou, Chen-Xi Ma, Ao-Guang Wang, Hui-Ben Zhang, Xiao-Hua Wu and Xiao-Ming Wang studied on Comparison and Improvement of Association Rule Mining Algorithm in 2015. They stated that In recent years, the data mining technology has been developed rapidly. New efficient algorithms are emerging. Association data mining plays an important role in data mining, and the frequent item sets are the highest and the most costly. This paper is based on the association rules data mining technology. The advantages and disadvantages of A-Priori algorithm and FP-growth algorithm are deeply analyzed in the association rules, and a new algorithm is proposed, finally, the performance of the algorithm is compared with the experimental results. It provides a reference for the extension and improvement of the algorithm of association rule mining.[5] Pooja R. Gaikwad, Shailesh D. Kamble, Nilesh Singh V. Thakur and Akshay S. Patharkar studied on Evaluation of A-Priori Algorithm on Retail Market Transactional Database to get Frequent Itemsets in 2017. They stated that In Data mining the concept of association rule mining (ARM) is used to identify the frequent item-sets from large datasets. It defines frequent pattern mining from larger datasets using A-priori algorithm & FP-growth algorithm. The A-Priori algorithm is a classic traditional algorithm for the mining all frequent item-sets and association rules. But, the traditional A-Priori algorithm have some limitations i.e. there are more candidate sets generation & huge memory consumption, etc. Still, there is a scope for improvement to modify the existing A-Priori algorithm for identifying frequent item-sets with a focus on reducing the computational time and memory space. This paper presents the analysis of existing A-Priori algorithms and results of the traditional A-Priori algorithm. Experimentation carried out on transactional database i.e. retail market for getting frequent itemsets. The traditional A-Priori algorithm is evaluated in terms of support and confidence of transactional itemsets.[6] Moushumi Sharma, Ajit Das and Nibedita Roy studied on A Complete Survey on Association Rule Mining and its Improvement in 2016. They stated that In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. Here they have classified Association rule mining in two ways, mining with candidate generation (A-Priori Algorithm) and mining without candidate generation (FP-Tree). Further they have classified these two algorithms into different phases. Based on the limitations of these algorithms different researchers gave different ways to improve the efficiency of these algorithms. In this paper they present a survey of some research work carried by different researchers .They hope that it will provide a guideline for the researchers in interesting research directions that have yet to be explored.[7] Mohammed Al-Maolegi and Bassam Arkok studied on an improved A-Priori algorithm for association rules in 2014. They stated that there are several mining algorithms of association rules. One of the most popular algorithms is A-priori that is used to extract frequent itemsets from large database and getting the association rule for discovering the knowledge. Based on this algorithm, this paper indicates the limitation of the original A-priori

algorithm of wasting time for scanning the whole database searching on the frequent itemsets, and presents an improvement on A-priori by reducing that wasted time depending on scanning only some transactions. The paper shows by experimental results with several groups of transactions, and with several values of minimum support that applied on the original A-priori and our implemented improved A-Priori that our improved A-Priori reduces the time consumed by 67.38% in comparison with the original A-Priori, and makes the A-Priori algorithm more efficient and less time consuming.

2.1 Research Gaps

• All the algorithms produce frequent item-sets on the basis of minimum support. A-priori algorithm is quite successful for market based analysis in which transactions are large but frequent items generated is small in number.

• The A-priori variations (DHP, DIC, Partition, and Sample) algorithms among them DHP tries to reduce candidate item-sets and others try to reduce database scan.

• DHP works well at early stages and performance deteriorates in later stages and also results in I/O overhead.
• For DIC, Partition, sample algorithm performs worse where database scan required is less then generating candidates.

• Vertical Layout based algorithms claims to be faster than A-priori but require larger memory space then horizontal layout based because they needs to load candidate, database and TID list in main memory.

• For projected layout based algorithms include FP-Tree and H-mine, performs better then all discussed above because of no generation of candidate sets but the pointes needed to store in memory require large memory space.

• FP-Tree variations include COFI-Tree and CT-PRO performs better than classical FP-tree as COFI-tree performs better in dense datasets but with low support its performance degrades for sparse datasets and for CT-PRO algorithm performs better for sparse as well for dense data sets but difficult to manage the compress structure. Therefore these algorithms are not sufficient for mining the frequent item-sets for large transactional database.

III. RESEARCH INVESTIGATION & METHODOLOGY

This research used to implement the proposed solution of the problem that is being taken care of in this thesis work, the following methodology is used:

- To analyze the various existing techniques and find their strengths and weakness by the literature survey.
- To compare the existing techniques.
- Build a program for our desired problem by using maximal A-Priori (Improved A-Priori technique) and FP-tree structure.
- Validate the program by desired input.

3.1 The Proposed Objectives

- The problems or the limitations defined in the above section of this chapter are proposed to be solved by:
- To observe the effect of various existing algorithms for mining frequent item-sets on various datasets.
- To propose a new scheme for mining the frequent item-sets for retailer transactional database i.e. for the above problem.
- To validate the new scheme on dataset.

3.2 Methodology

This thesis is conducted through review of the current status and the relevant work in the area of data mining in general and in the area of association rules in particular; analyze these works in the area of mining frequent item-sets; propose the new scheme for extracting the frequent item-sets based on hybrid approach of maximal A-priori and FP-tree algorithm that has high efficiency in term of the time and the space; validate its efficiency and seek avenues for further research. As frequent data item-sets mining are very important in mining the Association rules. Therefore there are various techniques are proposed for generating frequent item-sets so that association rules are mined efficiently.

General steps:
1. In the first pass, the support of each individual item is counted, and the large ones are determined.

2. In each subsequent pass, the large item-sets determined in the previous pass is used to generate new item-sets called candidate item-sets.

3. The support of each candidate item-set is counted, and the large ones are determined.

4. This process continues until no new large item-sets are found.

IV. EXPECTED OUTCOME

To perform the A-Priori and FP growth on the plate from with MATLAB. To find the support value, confidence and lift value of each inserted data base.

REFERENCE

[1] María N. Moreno, Saddys Segrera, Vivian F. López and M. José Polo, “Improving the quality of association rules by preprocessing numerical data”, Universidad de Salamanca, Plaza Merced S/N, 37008, Salamanca.
[2] Trupti A. Kumbhare and Prof. Santosh V.Chobe, “An Overview of Association Rule Mining Algorithms ”, Vol. 5 (1), 2014.
[3] Zainab Darwish, Mousa Al-Akhras and Mohamed Habib, “Use Filtering Techniques to Improve The Accuracy of Association Rules”, 2017 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT).
[4] Xiao-Feng Gu, Xiao-Juan Hou, Chen-Xi Ma, Ao-Guang Wang, Hui-Ben Zhang, Xiao-Hua Wu and

Xiao-Ming Wang, “Comparison and Improvement Of Association Rule Mining Algorithm”, 978-1-4673-8266-3/15/$3l.00 ©l015 IEEE.

[5] Pooja R. Gaikwad, Shailesh D. Kamble, Nileshsingh V. Thakur and Akshay S. Patharkar, “Evaluation of A-Priori Algorithm on Retail Market Transactional Database to get Frequent Item-sets”, pp. 187–192, Vol. 10, ISSN 2300-5963, 2017.

[6] Moushumi Sharma, Ajit Das and Nibedita Roy, “A Complete Survey on Association Rule Mining and Its Improvement”, Vol. 4, Issue 5, May 2016.

[7] Mohammed Al-Maolegi and Bassam Arkok, “An improved a-priori algorithm for association rules”, international journal on natural language computing, Vol. 3, No.1, February 2014.

[8] Amaranatha Reddy P, Pradeep G and Sravani M, “binary decision tree for association rules mining in incremental databases”, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.6, November 2015.

[9] Christian Borgelt, “Frequent item set mining”, Volume 2, November/ December 2012.