Optimizing Association Rule using Genetic Algorithm and Data Sampling Approach

Devyani Ojha
Department of computer Science and Engineering
Medicaps University, Indore, India

Pragya Pandey
Department of computer Science and Engineering
Medicaps University, Indore, India

ABSTRACT

In this paper work the association rules are optimized in order to find most suitable rules from the number of association rule generation algorithms. In this context most frequently used association rule mining algorithms are targeted for study namely Apriori and FP-tree. Basically the association rules are developed using transactional datasets. Additionally the number of generated rules in Apriori is large enough; on the other hand the FP-tree algorithm generates a main tree and additional trees. Such kind of tree generate confuse the experimenter. Therefore in this work a concept is proposed by which the optimal rules from both the set of rules are selected for applications.

In this context two different concepts of the rule selection techniques are used first technique usages the sampling technique and second directly usage the outcomes of the Apriori and FP-Tree algorithm and make search from one algorithm's rule set to others. In order to perform the search genetic algorithm is used which is used for optimal solution selection. According to the results sampling based technique needs additional computational resources as compared to genetic algorithm based technique due to additional evaluation cycles. But both the algorithms are effectively capable to reduce the amount of rules generated by the selected algorithms.

Keywords

Data Mining, Association Rule, FP-Tee, Apriori, Frequent Pattern mining, Association Rule Mining, Genetic Algorithm

1. INTRODUCTION

Data mining is a technique to modify, transform, calculate and extract the knowledgeable and fruitful data for applications. Therefore different kinds and nature of algorithms that are supported in data mining for extracting and evaluating the data patterns, among these algorithms the classification, clustering and association pattern mining are the popular techniques that are frequently used for mining the data. In this presented work the association rules mining or frequent pattern mining is the main area of research and development. In this context the key aim is associated to minimize the rules by optimizing them using the different techniques which are more useful or fit for specific amount or available transaction sets [1] [2].

A. Association Rule Mining

Data Mining is the discovery of hidden information found in databases and can be viewed as a step in the knowledge discovery process. Data mining functions include clustering, classification, prediction, and link analysis (associations). One of the most important data mining applications is that of mining association rules. An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent. Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the if/then statements have been found to be true [3].

Association rule mining has been an active research area in data mining, for which many algorithms have been developed. In data mining, association rule learning is a popular and well-accepted method for discovering interesting relations between variables in large databases. Association rules are employed today in many areas including web usage mining, intrusion detection and bioinformatics [4].

B. Frequent Item-set Mining

Frequent item set mining is one of the best known and most popular data mining methods. Originally developed for market basket analysis, it is used nowadays for almost any task that requires discovering regularities between (nominal) variables. Frequent sets play an essential role in many Data Mining tasks that try to find interesting patterns from databases, such as association rules, correlations, sequences, episodes, classifiers, clusters and many more of which the mining of association rules [5]. The identification of sets of items, products, symptoms, characteristics, and so forth that often occur together in the given database can be seen as one of the most basic tasks in Data Mining.

Frequent pattern mining has been an important subject matter in data mining from many years. A remarkable progress in this field has been made and lots of efficient algorithms have been designed to search frequent patterns in a transactional database. Frequent pattern mining can be used in a variety of real world applications. It can be used in super markets for selling, product placement on shelves, for promotion rules and in text searching [6] [7].

2. PROPOSED SYSTEM

The proposed modification on the association rules development is described in this chapter. Therefore initially the overview of the proposed work is provided and then detailed methodology of work is described. Finally the proposed technique is summarized using the algorithm steps.

A. Proposed Methodology

In order to optimize the association rules generated from Apriori and FP-Tree algorithm the following two suggestions are considered:

- ✓ Use of Boosting Technique
- ✓ Use of Optimization Algorithm

Both the techniques are designed and implemented, the discussion of both the techniques is provided as:

Boosting technique

The proposed boosting based technique for optimal rule selection technique is provided in figure 1. This figure involves the components that are used for processing of data to generate required set of rules:

Input dataset: any machine learning or data mining technique requires the initial data for evaluating the patterns from the data. These patterns are further used for prediction, classification, recommendation and other significant tasks. In this work the association rule computation is key task. Thus system needs to involve a transactional dataset as input. User provides a list of transactions as input for process and preparation of association rules.

Number of folds: that is second essential input for the system, user have to decide the number of folds under which the experiments are performed. That input is also used for computing the number of transactions in each generated new sample datasets. Let the dataset contains N number of total dataset patterns or transactions. Additionally the F number of sampling is required for experimentation. In this context the length of each sample L is calculated using the following formula:

$$L = \frac{N}{F}$$

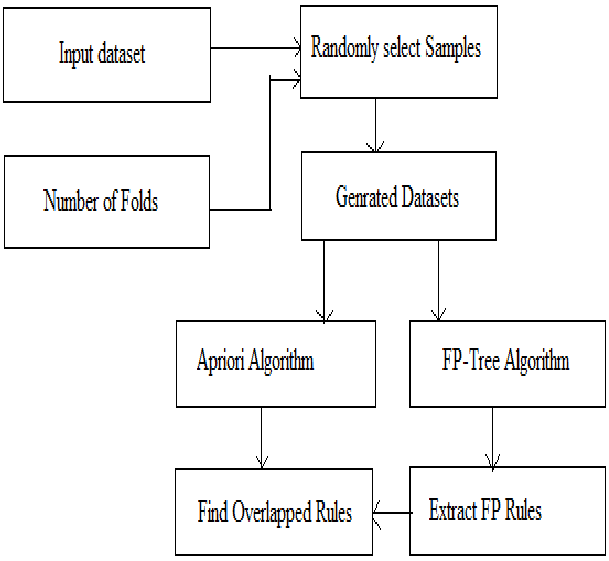

Figure 1: Boosting Based Technique

Randomly selected samples: in the previous phase inputs the following data is obtained:

1. total dataset transactions
2. number of sample datasets
3. length of sample in each sample dataset

Generated datasets: Using these data the system selects the random samples to generate the similar length and similar amount of dataset samples. These samples are used for next phase both the algorithms are applied on similar amount and randomly selected data samples for generating the association rules.

Apriori algorithm: the generated datasets are produced to the traditional Apriori algorithm for computing the Apriori algorithm that helps to generate association rules. Each the generated samples are evaluated individually for producing the different set of rules for each sample produced from previous step.

FP-tree algorithm: similar to the Apriori algorithm here the traditional FP-tree algorithm is also implemented for generating the frequent pattern trees. After generation of all the possible trees and sub-trees the outcomes are produced in next phase.

Extract FP-Rules: The output of FP-tree algorithm is accepted in this phase. That phase is responsible for generating the list of rules form the trees generated from last phase.

Find overlapped rules: In this phase both the kinds of generated rule sets are accepted from the Apriori and FP-tree algorithm. Both the set of rules are compared each set of rules and the 80% similar association rules which belongs to both the algorithm generated rules are selected as final overlapped rules.

Optimization Technique

That is the simple process followed for optimizing the list of rules from both the algorithms. Therefore the figure 2 is prepared to demonstrate how this process is working. The component description of the designed system is given as:

Input dataset: As described in previous system design the similar kind of transactional dataset is produced as input to the system.

Item sets: The entire dataset is evaluated in this phase for recovering the list of symbols available in dataset. This list of symbols is termed here as the item set for the algorithms.

Transaction set: The items recovered from the previous phase are responsible for generating the list of transactions. The available transactions are used separately for process with the algorithm.

Apriori algorithm: In this phase again the traditional Apriori algorithm is implemented which process the item sets and transaction sets for generating the list of association rules.

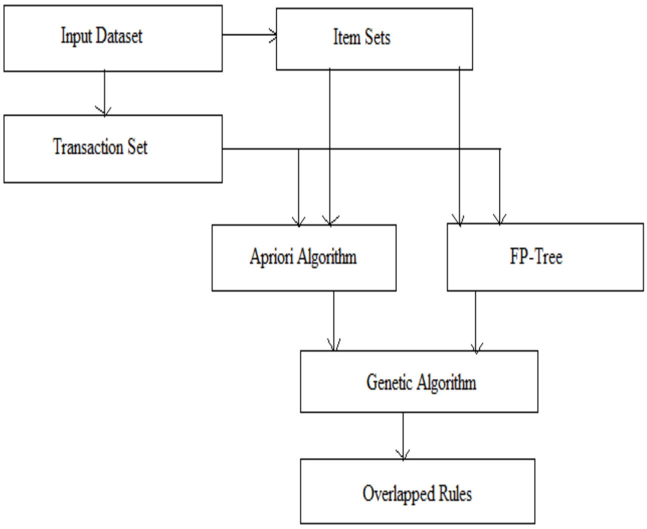

Figure 2: Genetic Algorithm Based

FP-Tree: In this phase the traditional FP-tree is used for generating the frequent pattern trees using the algorithm. In addition of that these trees are reformed as the association rules.

Genetic algorithm: The genetic search algorithm is implemented in this phase. This algorithm select one rule at a time from the FP-tree rules and Apriori algorithm based rules are used as the solution space for perform search for the similar kind of rule. Most fit rule from the list of Apriori based rules are selected as the optimum overlapped rule.

B. Proposed Algorithm

This section includes the algorithm steps using table 1 and table 2 of both the systems which demonstrated in previous section:

Table 1: Boosting Based Technique

```
Input: input dataset D, number of Folds F

Output: list of overlapped rules $L_O$

Process:

1.  $L_N = readDataset(D)$
2.  $S = \frac{N}{F}$
3.  $[Sam_1, Sam_2, \dots Sam_F] = SelectRandom(L_N, S)$
4.  $for(i = 1; i < F; i++)$
        a.  $AprioriRuleSet_i = Apriori.GenrateRules(Sam_i)$
        b.  $FPRuleSet_i = FPTree.GenrateRules(Sam_i)$
5.  $end\ for$
6.  $List_{ARS} = Combine(AprioriRuleSet_F)$
7.  $List_{FPR} = Combine(FPRuleSet_F)$
8.  $for(j = 1; j \leq List_{ARS}.length; j++)$
        a.  $for(k = 1; k \leq List_{FPR}.length; k++)$
                i.   $D = ComputeDisatnce(List_{ARS}^{i}, Li$
                ii.  $if(D \geq 0.75)$
                         1.  $L_O.Add(List_{FPR}^{k})$
                iii. $end\ if$
        b.  End for
9.  End for
10. Return $L_O$
```

Table 2: Genetic Algorithm Based Technique

```
Input: Dataset D

Output: List of Overlapped rules $L_O$

Process:

1.  $R = ReadDataset(D)$
2.  $List_{AR} = Apriori.genrateRules(R)$
3.  $List_{FPT} = FPTree.genrateRules(R)$
4.  $for(i = 1; i \leq List_{AR}.length; i++)$
        a.  $Temp = List_{FPT}^{i}$
        b.  $Result = GeneticAlgo.search(R, Temp)$
        c.  $L_o.add(Result)$
5.  End for
6.  Return $L_o$
```

3. RESULT ANALYSIS

A. Memory Consumption

The amount of main memory required to perform data analysis using the algorithm is termed here as memory usages or space complexity. The estimated comparative memory consumption of the implemented data mining algorithm is reported using figure 3 and table 3. Following are the formula by which we can estimate consumed memory:

$$\text{Memory Consumption} = \text{Total Memory} - \text{Free Memory}$$

Table 3 Tabular form of Memory Consumption

Number of Experiments	Proposed Approach	Traditional Approach
1	70.4707	103.2231
2	82.3261	115.6921
3	80.3362	121.6941
4	93.1123	119.3369
5	91.2256	133.0021
6	98.2561	125.7902
7	98.6981	138.0216

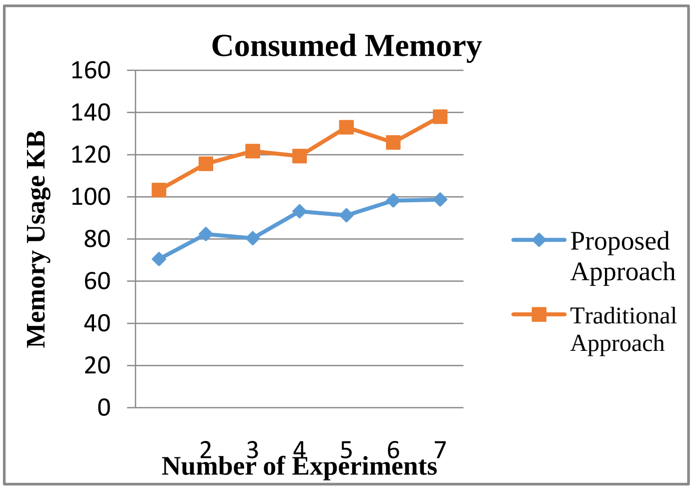

Figure 3 Memory Consumption

The memory consumption or space complexity of the implemented hybrid approach of FP-Tree and Apriori algorithm which is generated feasible solution using genetic algorithm is reported with the help of figure 3. In this diagram the X axis contains different experiments performed with system and Y axis shows amount of main memory consumed in terms of KB (kilobytes). By depiction of the given graph it clearly show that rule generation using genetic algorithm of hybrid approach is consuming less memory compared to only of FP-tree and apriori data mining traditional algorithm

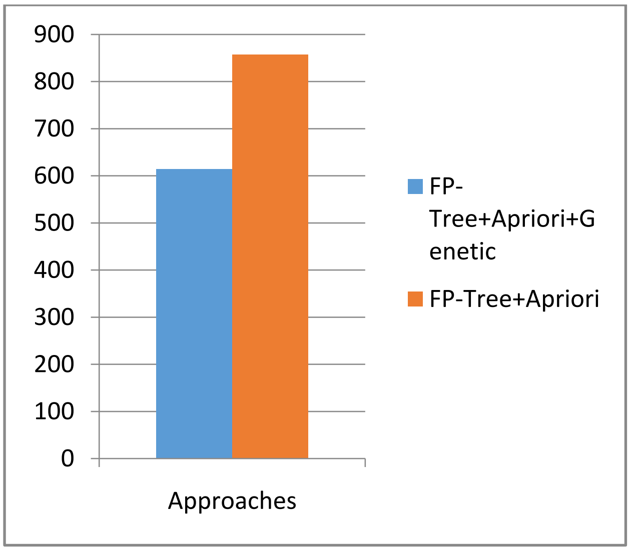

Figure 4 Memory Usage Mean Performance

Additionally to represent the performance of algorithms blue line shows the performance of hybrid approach of genetic algorithm where. Apriori and FP-tree based approach is denoted using orange line. In most of the experiments the performance of algorithms are fluctuating but it remains adaptable for data analysis in both the cases. In experimentations size of data is increases and their memory consumption is evaluated. During the experimentations that are observed if the number of candidate set generation is large then the memory requirement is higher and otherwise it remains fixed not much fluctuating. Furthermore, figure 4 show mean performance of the both implemented algorithm. FP-tree is much adoptable to apriori of efficient association rule mining.

B. *Time Consumption*

The amount of time consumed for processing the FP-tree and Apriori based association rules using the input datasets is termed here as the time consumption of algorithm or time complexity. Time consumption is to find below given formula:

$$\text{Time Consumed} = \text{End Time} - \text{Start Time}$$

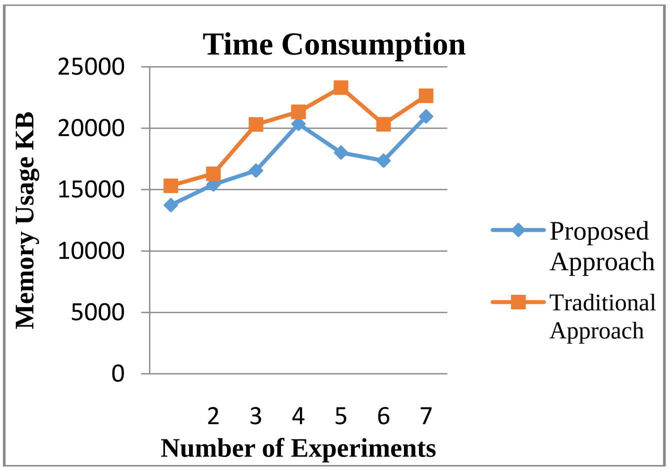

Figure 5 Time Consumption

The time consumption of the implemented approach is reported using the figure 5 and table 4. In this figure X axis shows the different experiments performed with the system and the Y axis shows the amount of time consumed for generate association rules in terms of milliseconds. According to the given performance the implemented genetic algorithm of rule generation consumes less amount of time as compared to the FP-tree and apriori based approach. But the time consumption is increases as the amount of data for association rule development is increases. Additionally, figure 6 depict mean value of both approaches. By this diagram, genetic based is much accurate and desirable that consumes less time whereas apiori + FP tree takes more time that is effective for rule mining

Table 4 Tabular form of time consumption

Number of Experiments	Proposed Approach	Traditional Approach
1	13747	15326
2	15423	16301
3	20316	20316
4	20361	21336
5	18022	23310
6	17362	20331
7	20691	22651

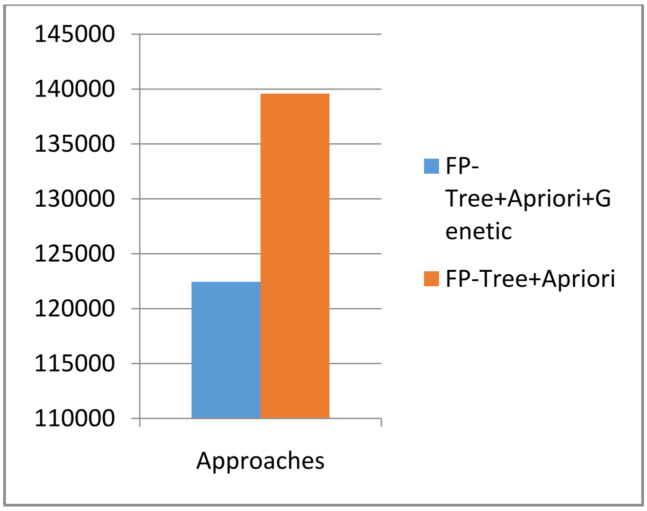

Figure 6 Time Consumption Mean Performance

4. CONCLUSION AND FUTURE WORK

The proposed work is aimed for improving the rules generated using traditional association rule mining techniques. In this context two popular approaches association rule mining technique namely Apriori and FP-tree is selected. In order to optimize the rule both algorithm utilized with the similar datasets and compared the obtained results. In first approach the boosting technique is used for improving the rule selection. In this context number of folds is accepted as input and the samples are selected randomly. These samples are used with the Apriori algorithm and FP-tree algorithm that results the set of rules. These set of rules are aggregated and the common rules are filtered as the optimum rules. In the second process in place of sampling of dataset the Meta heuristic algorithm is implemented for searching and optimizing the number of rules. Therefore the dataset is directly used with the traditional Apriori and FP-tree algorithm. Both the algorithms are generates rules according to their internal process. Now in next the FP-tree algorithm based rules are used as query for search and the Apriori algorithm based rules are used as solution space for search. Using the genetic search algorithm the fit solutions are collected as final outcome of the system.

The results obtained by evaluation of both the suggested and implemented approaches of association rule optimization is performs acceptably. But boosting based technique need additional time and memory resources due to additional cycles implemented. Thus both the technique provides optimal selection of rules but boosting based technique is cost effective as compared to genetic algorithm based technique.

5. REFERENCES

[1] Dhanalakshmi. D and Dr. J. Komala Lakshmi, “A Survey on Data Mining Research Trends”, A Survey on Data Mining Research Trends, Volume 3, Issue 10 October, 2014 Page No. 8911-8919

[2] Berson, Alex, and Stephen J. Smith, Building data mining applications for CRM, McGraw-Hill, Inc., 2002.

[3] Agrawal, Rakesh, and Ramakrishnan Srikant, "Fast algorithms for mining association rules." Proc. 20th international conference very large data bases, VLDB, Volume 1215, 1994.

[4] “Association Rules Mining”, available online at: https://www.vskills.in/certification/tutorial/data-mining-and-warehousing/association-rules-mining/

[5] Rana Ishita and Rana Ishita, “Frequent Itemset Mining in Data Mining: A Survey”, International Journal of Computer Applications (IJCA), Volume 139 – No.9, April 2016

[6] Sanjaydeep Singh Lodhi and Premnarayan Arya, “Frequent Itemset Mining Technique in Data Mining”, International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 5, July 2012

[7] Borgelt, Christian. "Frequent item set mining", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2.6 (2012): pp. 437-456.