
International Journal of Control Theory and Applications

ISSN : 0974-5572



Volume 10 • Number 4 • 2017

# Analysis of Association rules for Big Data using Apriori and Frequent Pattern Growth Techniques

Priyanka Chauhan[1], Chetna Dabas*[1] and J. P Gupta[2]

[1] *Department of Computer Science and Engineering, Jaypee Institute of Information Technology, JIIT, Noida, India*
[2]*Ex Chancellor, Lingaya's University, Faridabad, India*

***Abstract*:** In huge databases, the cost of finding association rules takes a toll on organizations. As the time advances, old transactions may be obsolete along with market-oriented applications and new transactions are formed. As a consequence, incremental updating techniques are required to be designed for the association maintenance and restrict redoing scanning of the entire database which is updated. The proposed implementation work for big data with Frequent Pattern Growth and Apriori techniques is carried out on WEKA 3.8.0. The comparison of these two algorithms for the considered dataset is also carried out as a part of this work and it is performed in terms of varying number of instances and attributes. It is concluded that the Frequent Pattern Algorithm performs better than the Apriori algorithm for the considered dataset in terms of execution time.

***Keywords:*** WEKA, Big Data, Frequent Pattern Growth, Apriori

## I. INTRODUCTION

In the current world, the collection of data is increasing day by day as it is important for the storage of database or dataset to keep the record. Storing and using the large data is not an issue, but getting the appropriate information from that data is quite a difficult job to do. The analysis of that collected data is made possible by many data mining techniques. In data mining we find the relation and patterns between the sets of items of larger relational databases which can help in predicting and improving the performance of the system. The relations between the data in data mining are found by a well-known approach, that is, association rule mining. Many association rules are found that relates the dependency of data on each other. Large number of association rules is generated by which we can also classify the kinds or class of database instances.

Association rule mining can define all the relationships even in moderate dataset. But the motive of association rule mining is not finding all the relationships but the set of interesting ones. The interestingness depends on the application. Therefore the set of rules are generated and are pruned to get rid of unnecessary association rules. Two strategically measures of association rule are support and confidence. These are the user defined measures of interestingness. The two terms support and confidence are the statistical significance of a rule and degree of certainty, respectively.

Association rule [1] is represented as the implication of X->Y which is an if-then relationship of X and Y, that is, how many times Y has occurred when X is already occurred and the association rules' interestingness depends on the values of support and confidence. Here X is an antecedent and Y is a consequent, or LHS (Left Hand Side) and RHS (Right Hand Side), respectively.

Support of X and Y= sup(X, Y)= probability that X and Y (XUY) both are there in the same transaction.

Confidence of X and Y= Conf(X=>Y)= P(Y|X)= Sup(XUY)/Sup(X)= the conditional probability that the transaction contains RHS under the condition that transaction also contains LHS.

To generate the rules, first it finds the frequent sets of item which are called itemsets. Then using these itemsets the rules are generated by applying the mining theorems which satisfy the minimum support and confidence value.

The rules generated give some specific knowledge which can be used in many application domains such as stock analysis, weather forecast, credit risk assessment, market basket analysis, medical diagnosis, etc.

The whole study is organized in the following manner: literature study related to association rule mining techniques us given in Section II, existing methods and techniques are discussed in Section III, Section IV presents the description of data set used and results analysis and finally the entire study is concluded in section V.

## II. LITERATURE REVIEW

In paper [2], author has used the FP-Growth algorithm for the classification of the cancerous masses into two categories of benign and malignant. Association rules can help the doctors in decision making and medical diagnosis on the basis of relation of tests performed for the particular disease. Breast cancer Wisconsin Dataset of 699 instances and 10 attributes has been used for the extraction of association rules which provides high accuracy.

The use of association rule mining play an important role for the analysis of road accidents in India. As discussed in paper [3] and [4], apriori algorithm is applied to the road accident data set and the causes of accident are considered as the attributes. The large data set is classified into number of clusters and then the association rule mining techniques are applied to them to generate more efficient rules. It can help to reduce accident happening, find main factor and circumstances of causing accidents so that we can try to avoid them.

In [5] paper, the author have proposed an new algorithm for the incremental dataset, that is, the algorithm to be applied on dymanic dataset with some constraints on them so that when the number of transaction data in increased the older rules don't get void and number of the generation of new rules is very less that it doesn't affect the whole relationship of association. In this paper the constraint association rules are constructed in the form of a tree then the association rule mining algorithm is applied in an iterative manner to get the effective rules for the incremental dataset. This proposed algorithm improves the accuracy for the constraint rules. But there may be possibility that the duplicate rule were not filtered out so this needs to be done in future.

In [2], we studied that FP-Growth algorithm was used to analyze and classify the cancerous masses. Now in [6] the author proposed and algorithm to classify the brain tumor MRI images by applying the association rule mining first, i.e., apriori algorithm, and then optimizing those rules using Multi Objective Genetic Algorithm. The 3D images are converted into 2D images and these 2D images are further divided into fragments and algorithm is applied to them. The proposed algorithm gives better accuracy than the original one.

The best example of association rule mining is the market basket analysis. Same as this example, in [7], a retail company called XMART implemented the association rules with apriori algorithm on its transaction data set to know the pattern in which the customers buy products so that the products can be arranged in a way that all

the related products are in a same block. But the company has 8 different stores at 8 different places so the data of all the stores is divided into eight clusters and the association rules mining is applied to each cluster to generate the rules which can be used to increase the quality of promotion. But as there are 8 clusters so it is not necessary that all the rules of one store are applicable to other stores means a rule could be applied to one store but could not be applied to other store.

Association rule mining helps in decision making, so in [8] association rules have been used to make the strategy for the IIC World Cup 2015 based on the earlier dataset of scores, strike-rates etc. of all the Indian team players. The performance of all the batsman and bowlers are analyzed by generating rule so that the best players can be chosen for the team to play and get the best result. In this paper, apriori algorithm has been applied on two datasets of all the ODIs which are Dataset-1 for bowlers and Dataset-2 for batsmen. However this approach proved to be helpful to planning team strategy but there are some limitations. The condition of ground, weather conditions etc. have not been considered so there remains a hope for the improvement in this approach.

This paper [9] uses the apriori algorithm and optimizes it for the time-memory domain. This proposed approach is divided into two steps. It splits the apriori algorithm into two phases. The first phase is same as calculating all the itemsets in every transaction with their frequencies that too without pruning and the second phase is producing association rules using the itemsets produced in phase one and their frequencies. The proposed approach was designed to work on a server-client based system and it can also be applied to a single computer system. It reduces the fetching time of each transaction to only once.

## III. EXISTING METHODS AND TECHNIQUES

Many different algorithms and techniques for association rules generation were presented over time. Some of the well-known algorithms are Apriori, FP-Growth, ECLAT, Predictive Apriori and many more. Here we are describing these techniques, and will further demonstrate FP-Growth and apriori algorithms using the WEKA tool.

### Apriori Algorithm

The apriori algorithm is used for mining frequent itemsets for boolean association rules. This algorithm is proposed by Agrawal R in 1994 [11]. The name of the algorithm is based on the fact that it uses the prior knowledge of the frequent item set properties. It is designed to operate on database containing transactions of the items. Apriori uses bottom up approach in which frequent subsets are extended one item at a time. It is an iterative approach and each iteration contains two steps. First one is generation of candidates from set of items and second step is pruning the generated candidate to eliminate the infrequent item sets. The process of this algorithm is given in figure 1. It uses BFS strategy to count the support of items. This algorithm is easy to implement and is parallelized but its disadvantages is that it need multiple scans of database which increases the execution time and memory space.

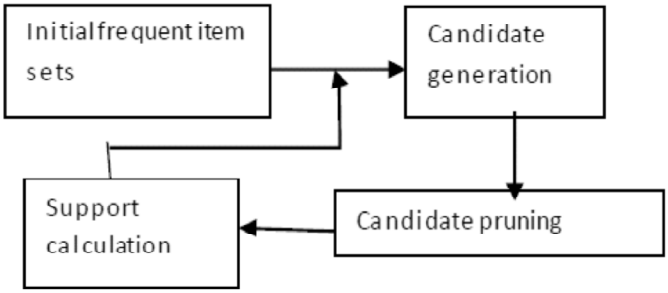


**Figure 1: Process of Apriori Algorithm**

The pseudo code for this algorithm is given in block procedure Apriori() as follows:

```
procedure Apriori (T, minSupport) { //T is the database and minSupport is the minimum support
        L1= {frequent items};
        for (k= 2; Lk-1 !=Ø; k++) {
                Ck= candidates generated from Lk-1
                //that iscartesian product Lk-1 x Lk-1 and eliminating any k-1 size itemset that is not
                //frequent
                for each transaction t in database do{
                        #increment the count of all candidates in Ck that are contained in t
                        Lk = candidates in Ck with minSupport
                }//end for each
        }//end for
        return ∪k Lk;
}
```

FP-Growth Algorithm

Frequent Pattern Growth (FP-Growth) algorithm, proposed by Han in 2000 [12], uses an extended prefix tree FP tree structure to store the database in a compressed form. It adopts the divide and conquers strategy. It is an efficient and scalable algorithm for finding the relation between the itemsets using the pattern fragment growth. In this algorithm ther are two processes to generate the association rules, i.e., construction of prefix tree and then using this prefix tree (called FP-Tree) the FP-Growth algorithm further moves on to generate association rules. The FP-Growth algorithm with FP-Tree is given in block Procedure FP-growth() as follows:

```
call FP-growth(FP-tree, null).
Procedure FP-growth (Tree, A)
{
if Tree contains a single path P
then for each combination (denoted as B) of the nodes in
the path P do
generate pattern B □ A with support=minimum support of
nodes in B
else for each ai in the header of the Tree do
{
generate pattern B = ai □ A with support = ai.support;
construct B's conditional pattern base and B's conditional
FP-tree
TreeB;
if TreeB ≠ Ø
then call FP-growth (TreeB, B)
}
}
```

The comparison based on the whole literature review and the features of both the algorithms is discussed in Table I.

## IV. EXPERIMENTAL SETUP

There are many tools and software for mining the data such as R, MEXL, SAS, XLMINER etc., so we analyzed the data set in tool called WEKA 3.8.0 which is a java based machine learning tool. We have used Weka Explorer which is a very useful interface for generating association rule form the transactional data. Experiment

**Table I**
**Comparison of Association Rule Mining Algorithms**

| *Algorithm Name* | *Applications* | *Merits/Demerits* | *Accuracy* | *Data Support* |
|---|---|---|---|---|
| **Apriori** | Best for closed item sets. | Slower, takes more memory, generates candidate sets, | Less | Limited |
| **FP-Growth** | Used in cases of large problems as it doesn't require generation of candidate sets. | Scan database only twice, faster, generates complex tree structure | More | Very large |

performs on following hardware configuration: RAM 4GB, Processor Intel Core i3. The dataset used is taken from UCI and Tunedit Machine Learning Repository. The work involves the database which has Boolean detail of the attribute in arff format i.e. supermarket.arff. There are three datasets used in this experiment as follows:

1. Supermarket.arff: It is the transactional dataset of customers for a super market which holds the details of products bought by customers and the departments involved, which consists of 4627 instances and 217 attributes.
2. Vote.arff: It consists of the voting records of the US House of Representatives which consists of 435 instances and 17 attributes.
3. Spect_Heart.arff: It consists of the cardiac diagnosing of the patients which have 187 instances and 23 attributes.

## V. WORK DONE

The datasets were collected from the machine learning repositories, the data should be in tabular form. The datasets collected are in form of ARFF (Attribute-Relation File Format) file format which is loaded into Weka software by using preprocess option in Weka Explorer, then the data is transformed from relational or other form of data to transactional dataset because the association rule mining algorithm can only be applied to transactional dataset in Weka. Now by using associate option we go to select the appropriate algorithm we want to implement on dataset to generate the association rule and set the value for different metrics such as support and confidence and run the algorithm. The best association rules are then generated as output. The rules are displayed in "Association output" box. The flowchart for the implementation of mining techniques is shown in figure 2.

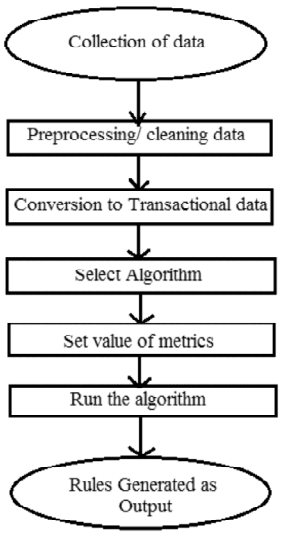


**Figure 2: Flowchart of Implementation of Association Rule Mining Technique**

## VI. RESULT ANALYSIS

The implementation of apriori and FP-Growth algorithm is shown below for the supermarket dataset which contains 4627 instances and 217 attributes. The algorithms are performed on other datasets also. The Association rules by apriori algorithm is shown in figure 3(a) and figure 3(b) and value for support and confidence and other properties are shown in figure 4. The Association rules by FP-Growth algorithm is shown in figure 5 and value for support and confidence and other properties are shown in figure 6.

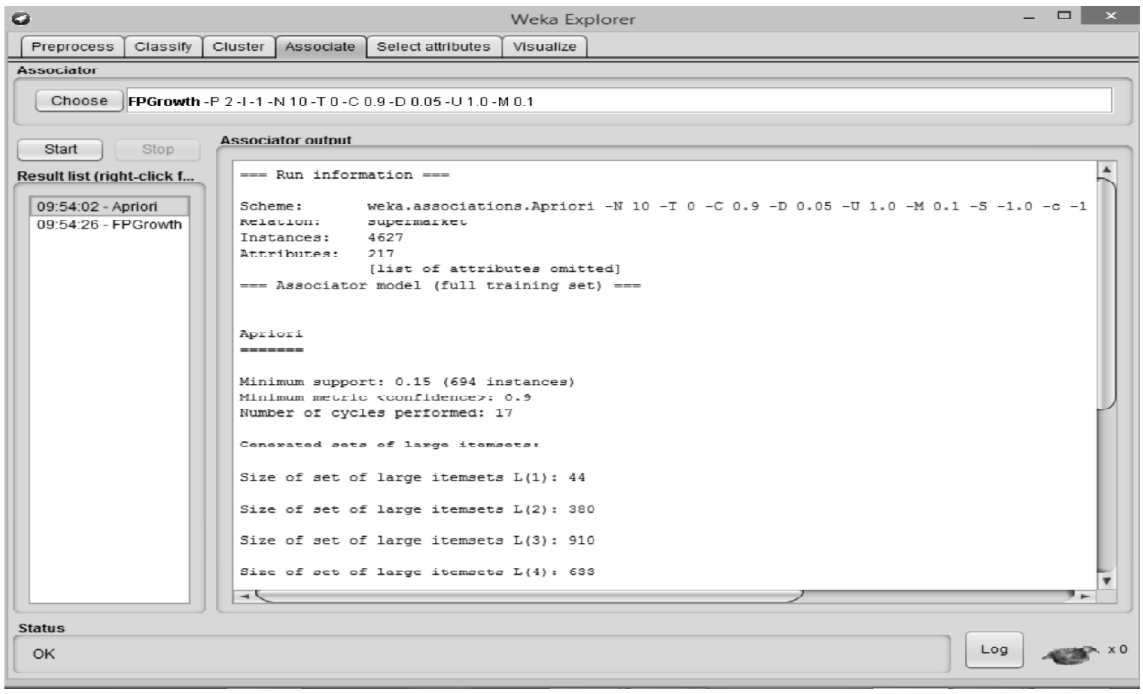


**Figure 3(a): Rules Generated by Apriori Algorithm**

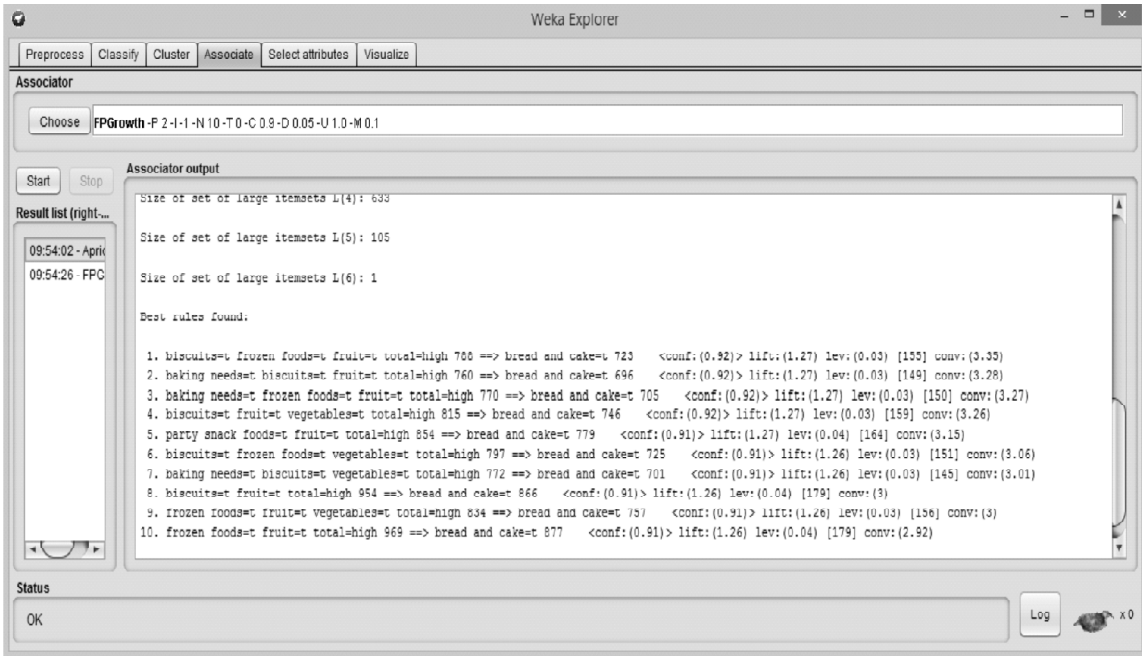


**Figure 3(b): Rules Generated by Apriori Algorithm**

The association rules mining algorithms are applied to the Boolean transactional datasets which we get after the conversion from relation databases. Here minimum support value is 0.1, minimum Confidence value is 0.9 and the number of rules to be generated is 10. We can change these values according to the need. But these values should be greater for the larger datasets otherwise the accuracy will be less for lesser values of these metrics.

The rules satisfying these minimum support and minimum Confidence values are the legal rules and are shown as the output.

One more metric is taken called lift. List is the ration of the total support to the expected support if both items were independent. It can be formulated as:

Lift(X=>Y)= sup(XUY)/[sup(X)*sup(Y)]

And the highest value obtained in this example is 1.27.

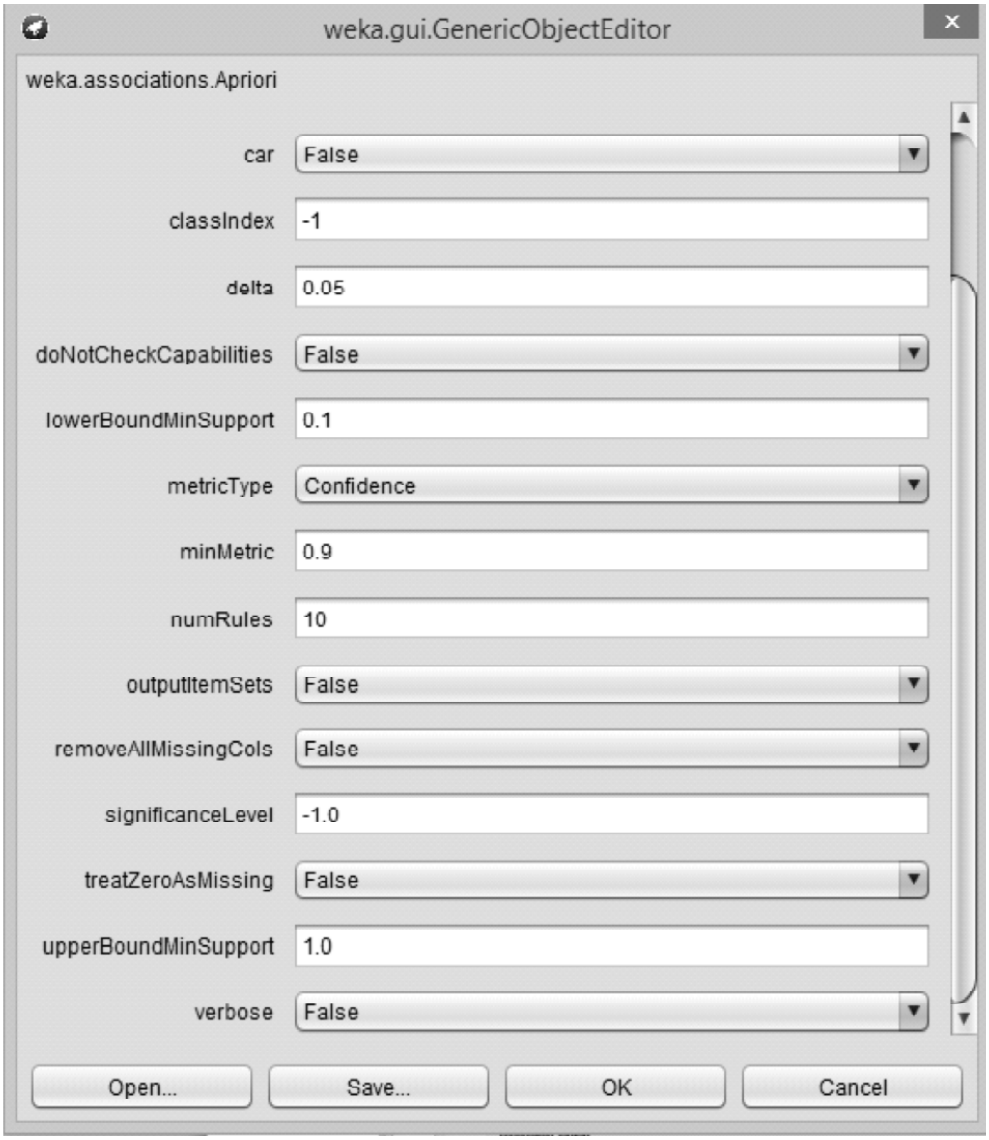


**Figure 4. Values for Apriori Algorithm**

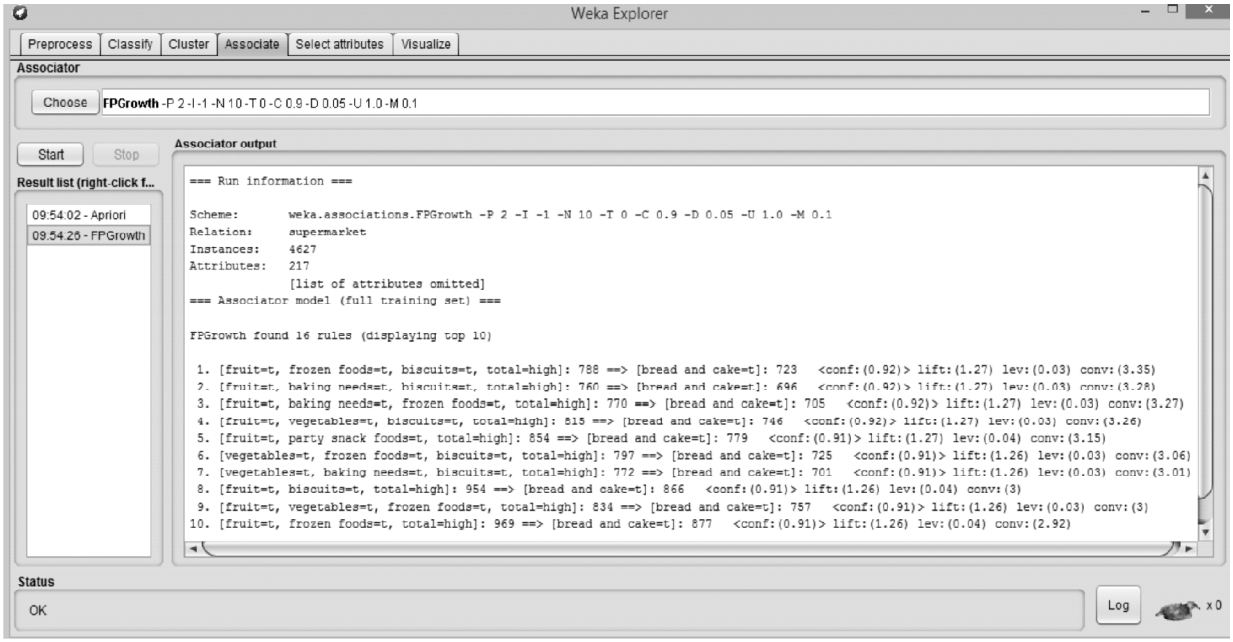


**Figure 5. Rules Generated by FP-Growth Algorithm**

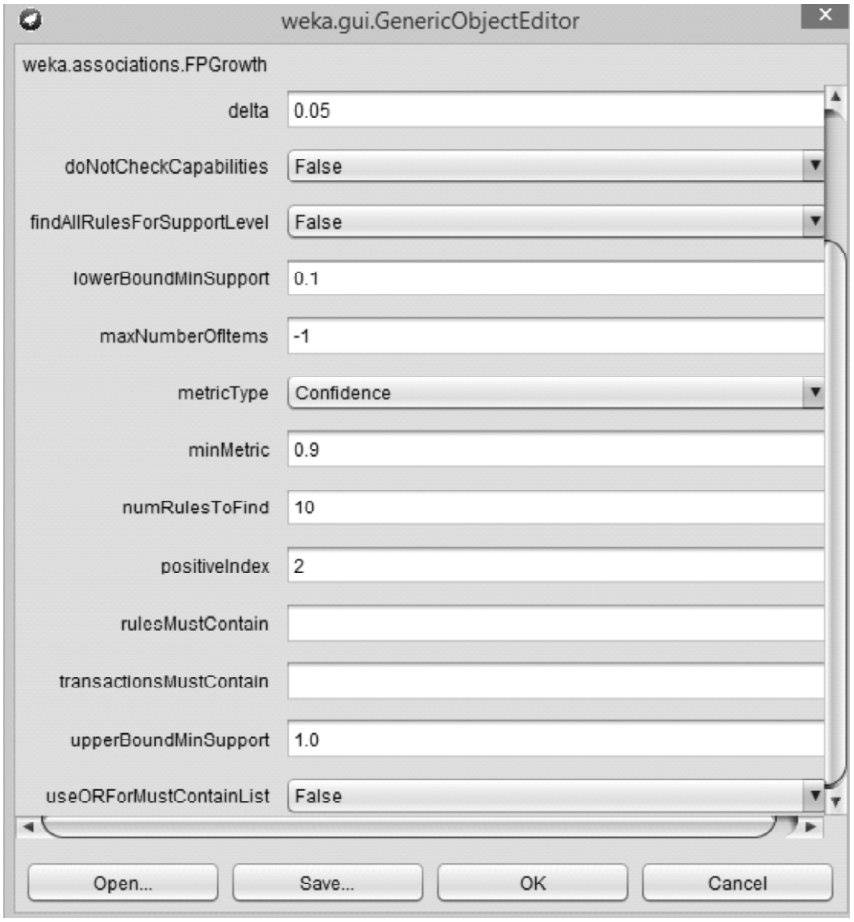


**Figure 6. Values for FP-Growth Algorithm**

The comparison of different data sets with different number of instances is given in TABLE II. Here the TABLE II shows that there is a huge difference in execution time of both the algorithms for all the datasets used in this experiment. The time taken by apriori algorithm is much more than FP-Growth algorithm for the factor of different number of instances. Many other factors are also need to be considered.

**Table II**
**Execution Time for Different Number of Instances**

| *Data Set (No. of instances)* | *Execution Time (in seconds)* | |
|---|---|---|
| | *Apriori* | *FP-Growth* |
| Super Market (4627) | 19 | 5 |
| Vote (435) | 9 | 2 |
| Spect_Heart (187) | 3 | 1 |

## VII. CONCLUSION

Analysis of association rules plays an important role in data mining to get the result for future and predict the decision. The performance analysis of both Apriori and FP-Growth algorithm is done by applying these algorithms on different datasets with different number of instances. Here the performance and execution time of FP-Growth algorithm is far better than Apriori algorithm. FP-Growth overcomes the drawbacks of apriori. So we conclude that FP-Growth behaves better than Apriori algorithm. This study shows that many techniques have been applied to apriori algorithm to improve its performance and FP-Growth is not suitable for the incremental dataset so further research has to be done.

## REFERENCES

[1] Jiawei Han, Micheline Kamber: "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, Elsevier, pp. 243-262, 2012.

[2] Majid Khalilian, Seyedeh Talayeh Tabibi: "Breast Mass Association Rules Extraction to Detect Cancerous Masses", Second International Congress on Technology, Communication and Knowledge (ICTCK 2015), IEEE, November 2015.

[3] Suwarna Gothane, Dr. M. V. Sarode: "Analyzing Factors, Construction of Dataset, Estimating importance of factor and generation of association rules for Indian road Accident", IEEE 6th International Conference on Advanced Computing, IEEE, 2016.

[4] Sachin Kumar, Durga Toshniwal: "A data mining framework to analyze road accident data", Journal of Big Data, Springer Berlin Heidelberg, 2015.

[5] Subbulakshmi. B, Monisha. M: "Incremental Constraint Class Association Rule Mining of Student Performance Dataset", Fifth International Conference on Recent Trends in Information Technology (ICRTIT), IEEE, 2016.

[6] Poonam Sonar, Udhav Bhosle: "Optimized Association Rules for MRI Brain Tumor Classification", 3rd International Conference on Computing for Sustainable Global Development (INDIACom), IEEE, 2016.

[7] Wildan Budiawan Zulfikar, Agung Wahana, Wisnu Uriawan, Nur Lukman: "Implementation of association rules with apriori algorithm for increasing the quality of promotion", International Conference on Cyber and IT Service Management, IEEE, 2016.

[8] Sanmoy Bhattacherjee, Jayakrushna Sahoo, Adrijit Goswami: "Association Rule Mining Approach in Strategy Planning for Team India in ICC World Cup 2015", Second International Conference on Advances in Computing and Communication Engineering, IEEE, 2015

[9] Polla A. Fatah, Ibrahim Hamarash: "Optimization of Association Rule Mining-A Two Step Breakdown Variation of Apriori Algorithm", Internet Technologies and Applications (ITA), IEEE, 2015.

[10] Rakesh Agrawal and Ramakrishnan Srikant: "Fast algorithms for mining association rules in large databases", Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499, Santiago, Chile, September 1994.

[11] Jiawei Han, Jian Pei, Yiwen Yin: "Mining Frequent Patterns without Candidate Generation". Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. SIGMOD '00, Vol 29, June 2000.