International Journal of Advanced Technology in Engineering and Science
Vol. No.3, Special Issue No. 01, September 2015
www.ijates.com
ijates
ISSN 2348 - 7550

ASSOCIATION RULE MINING WITH APRIORI AND FPGROWTH USING WEKA

Ajay Kumar Mishra[1], Dr. Subhendu Kumar Pani[2],
Dr. Bikram Keshari Ratha[3]

[1]PhD Scholar, [3]Reader, Utkal University,Odisha, (India)
[2]Associate Prof., Dept. of CSE,OEC,BPUT, Odisha, (India)

ABSTRACT

Association rule mining is considered as a Major technique in data mining applications. It reveals all interesting relationships, called associations, in a potentially large database. However, how interesting a rule is depends on the problem a user wants to solve. Existing approaches employ different parameters to guide the search for interesting rules. Class association rules which combine association rule mining and classification are therefore concerned with finding rules that accurately predict a single target (class) variable. The key strength of association rule mining is that all interesting rules are found. The number of associations present in even moderate sized databases can be, however, very large – usually too large to be applied directly for classification purposes. Therefore, any classification learner using association rules has to perform three major steps: Mining a set of potentially accurate rules, evaluating and pruning rules, and classifying future instances using the found rule set. In this work, we make a comparison of association rule mining algorithms. We use two most popular algorithms namely Apriori and filtered Associator using SPECT heart dataset available Tunedit Machine Learning Repository .

Keywords: Association Rule Mining, Apriori, FPGrowth.

I. INTRODUCTION

Data mining is considered to be an emerging technology that has made revolutionary change in the information world. The term `data mining' (often called as knowledge discovery) refers to the process of analyzing data from different perspectives and summarizing it into useful information by means of a number of analytical tools and techniques, which in turn may be useful to increase the performance of a system. Technically, “data mining is the process of finding correlations or patterns among dozens of fields in large relational databases”. Therefore, data mining consists of major functional elements that transform data onto data warehouse, manage data in a multidimensional database, facilitates data access to information professionals or analysts, analyses data using application tools and techniques, and meaningfully presents data to provide useful information. According to the Gartner Group, ``data mining is the process of discovering meaningful new correlation patterns and trends by sifting through large amount of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques''[3] . Thus use of data mining technique has to be domain specific and depends on the area of application that requires a relevant as well as high quality data. More precisely, data mining refers to the process of analyzing data in order to determine patterns and their relationships. It automates and simplifies the overall statistical process, from data source(s) to model application. Practically analytical techniques used in data mining include statistical methods and mathematical modeling. However, data mining and knowledge discovery is a rapidly growing area of research and application that builds on techniques and theories from many fields, including statistics, databases, pattern recognition, data visualization, data warehousing and OLAP, optimization, and high performance computing [1,2] . Worthy to mention that online analytical processing (OLAP) is quite different from data mining, though it provides a very good view of what is happening but cannot predict what will happen in the future or why it is happening. In fact, blind applications of algorithms are not also data mining. In particular, ``data mining is a user-centric interactive process that leverages analysis technologies and computing power, or a group of techniques that find relationships that have not previously been discovered'' [4,6,7] . So, data mining can be considered as a convergence of three technologies -- viz. increased computing power, improved data collection and management tools, and enhanced statistical algorithms. Data and information have become major assets for most of the organizations. The success of any organization depends largely on the extent to which the data acquired from business operations is utilized.

Association rule mining is a widely-used approach in data mining. Association rules are capable of revealing all interesting relationships in a potentially large database. The abundance of information captured in the set of association rules can be used not only for describing the relationships in the database, but also for discriminating between different kinds or classes of database instances. However, a major problem in association rule mining is its complexity. Even for moderate sized databases it is intractable to find all the relationships. This is why a mining approach defines a interestingness measure to guide the search and prune the search space. Therefore, the result of an arbitrary association rule mining algorithm is not the set of all possible relationships, but the set of all interesting ones. The definition of the term interesting, however, depends on the application. The different interestingness measures and the large number of rules make it difficult to compare the output of different association rule mining algorithms. There is a lack of comparison measures for the quality of association rule mining algorithms and their interestingness measures. Association rule mining algorithms are often compared using time complexity. That is an important issue of the mining process, but the quality of the resulting rule set is ignored. On the other hand there are approaches to investigate the discriminating power of association rules and use them according to this to solve a classification problem [5,8]. This research area is called classification using association rules [9]. It has to deal with a large number of rules.

Therefore, rule selection and rule weighting are essential for these approaches in classification. An important aspect of classification using association rules is that it can provide quality measures for the output of the underlying mining process. The properties of the resulting classifier can be the base for comparisons between different association rule mining algorithms. A certain mining algorithm is preferable when the mined rule set forms a more accurate, compact and stable classifier in an efficient way. In the next section, we provide an overview of data mining concepts, its process, different techniques and their potential applications. In section 3, we describe our study on finding the best set of class association rules for higher predictive accuracy. Finally the paper concludes in section 4 with a glimpse to our future work.

II. TECHNIQUES AND ALGORITHMS

Classification approach can also be used An Empirical Study on Class Association Rules Mining for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality. Some commonly used clustering methods are: a) Partitioning Methods b) Hierarchical Agglomerative (divisive) methods c) Density based methods d) Grid-based methods e) Model-based methods.

2.1 Apriori

Application of the Apriori algorithm is a great achievement in the history of mining association rules[6].This technique uses the property that any subset of a large itemset must be a large itemset. Also, it is assumed that items within an itemset are kept in lexicographic order. The Apriori generates the candidate itemsets by joining the large itemsets of the previous pass and deleting those subsets which are small in the previous pass without considering the transactions in the database. By only considering large itemsets of the previous pass, the number of candidate large itemsets is significantly reduced.

2.2 FP Growth

The FP-Growth Algorithm, proposed by Han[11], is an efficient and scalable method for mining the complete set of frequent patterns by pattern fragment growth, using an extended prefix-tree structure for storing compressed and crucial information about frequent patterns named frequent-pattern tree (FP-tree). In his study, Han proved that his method outperforms other popular methods for mining frequent patterns, e.g. the Apriori Algorithm and the TreeProjection.It was proved that FP-Growth has better performance than other methods, including Eclat and Relim. The popularity and efficiency of FP-Growth Algorithm contributes with many studies that propose variations to improve his performance

III. EXPERIMENTAL STUDY AND ANALYSIS

3.1 WEKA Tool

We use WEKA (www.cs.waikato.ac.nz/ml/weka/), an open source data mining tool for our experiment. WEKA is developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA language. WEKA is a state-of-the-art tool for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data pre-processing, feature reduction, classification, regression, clustering, and association rules. It also includes visualization tools. The new machine learning algorithms can be used with it and existing algorithms can also be extended with this tool.

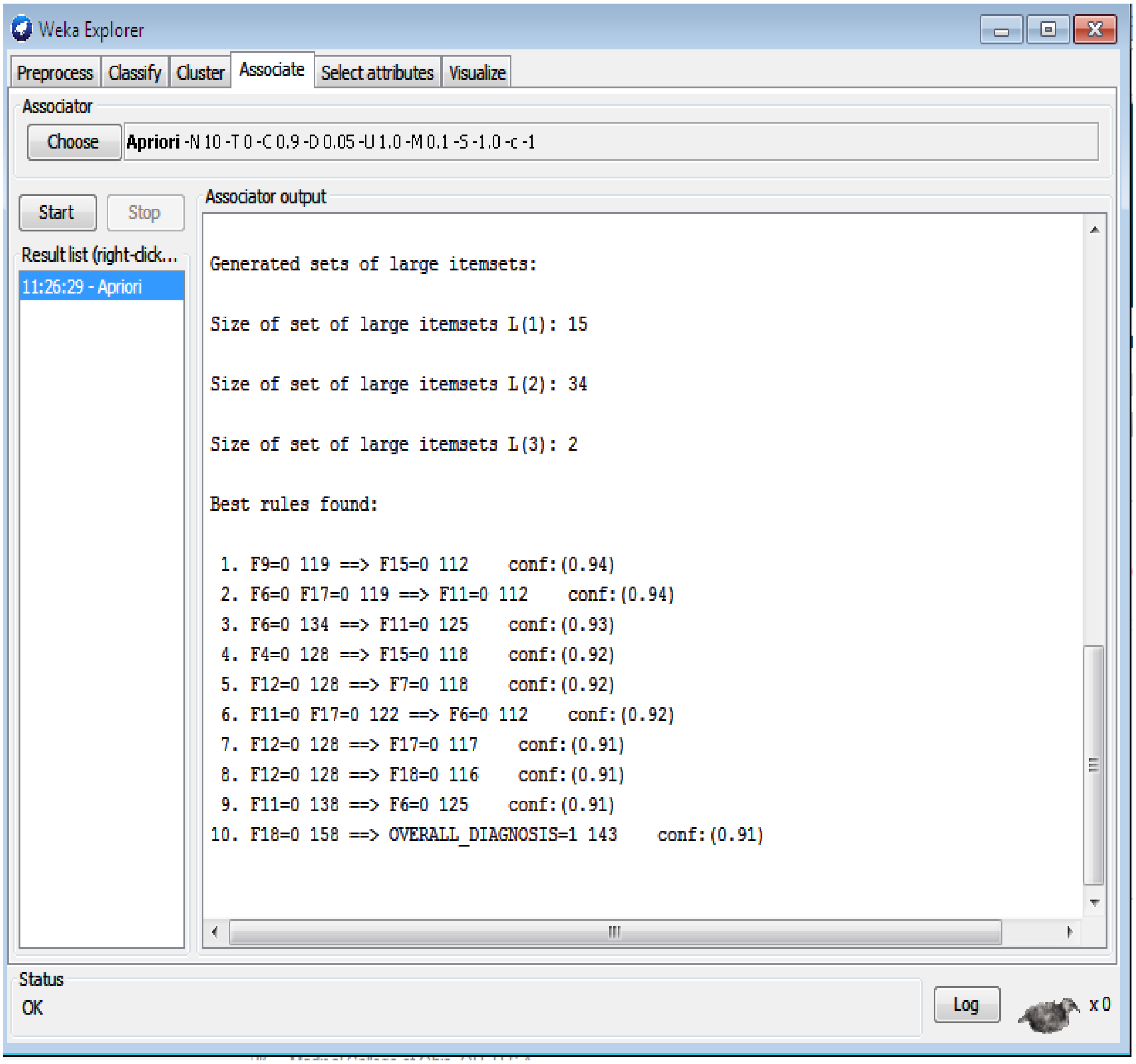
3.2 Dataset Description

We performed computer simulation on a SPECT heart dataset available Tunedit Machine Learning Repository [10]. It contains 187 instances and 23 attributes. The features describe different factor for heart diesease reoccurrence.. There are 2 instances having missing values.

3.3 Results Analysis

The class association rules generated by Apriori algorithm on the original dataset is given in Figure-1 and rules generated by FPGrowth is shown in Figure-3. Results are derived from the properties which is described in Fig2 and fig.4 respectively .

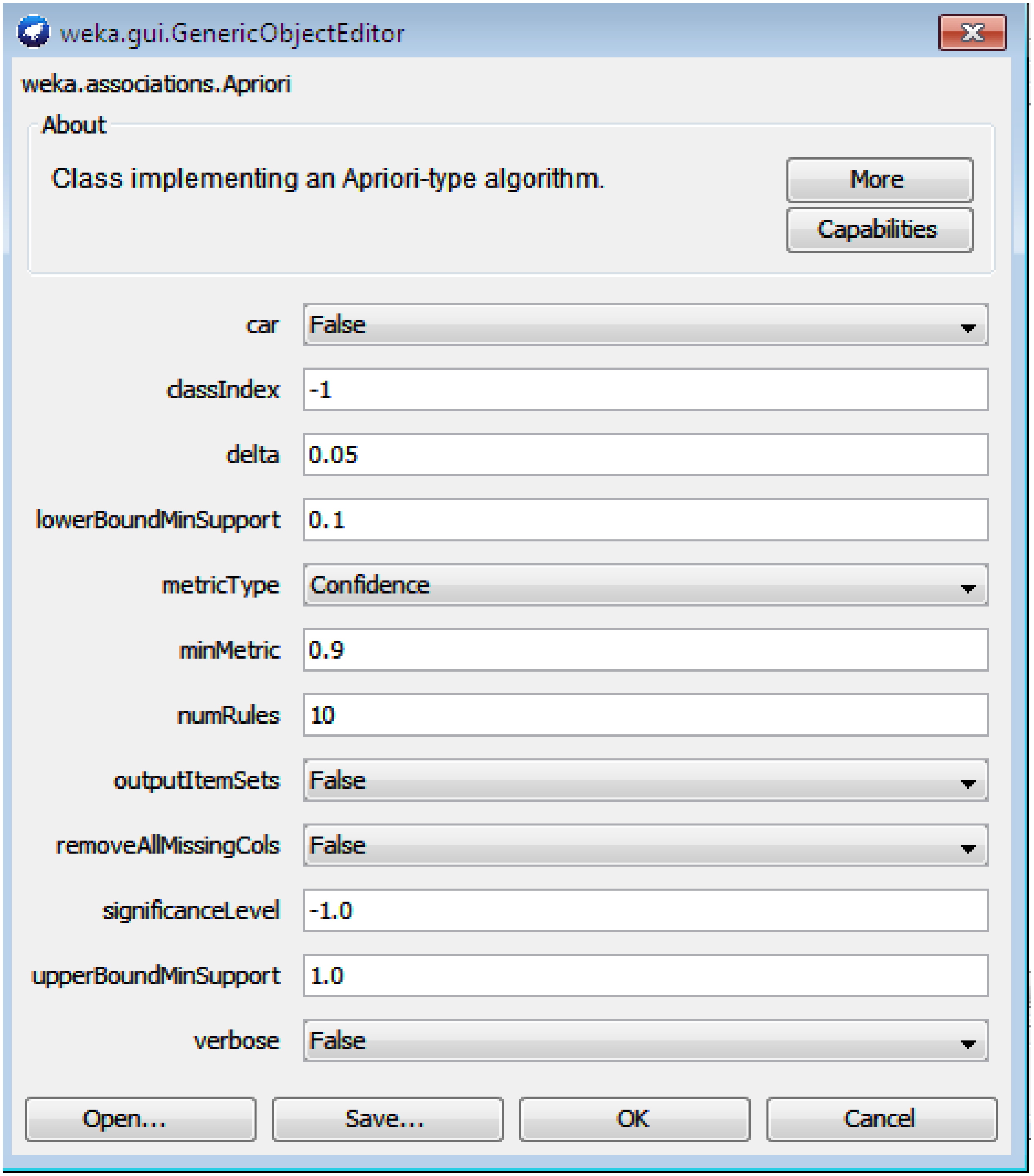
3.4 Apriori



Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

```
Generated sets of large itemsets:

Size of set of large itemsets L(1): 15

Size of set of large itemsets L(2): 34

Size of set of large itemsets L(3): 2

Best rules found:

 1. F9=0 119 ==> F15=0 112    conf:(0.94)
 2. F6=0 F17=0 119 ==> F11=0 112    conf:(0.94)
 3. F6=0 134 ==> F11=0 125    conf:(0.93)
 4. F4=0 128 ==> F15=0 118    conf:(0.92)
 5. F12=0 128 ==> F7=0 118    conf:(0.92)
 6. F11=0 F17=0 122 ==> F6=0 112    conf:(0.92)
 7. F12=0 128 ==> F17=0 117    conf:(0.91)
 8. F12=0 128 ==> F18=0 116    conf:(0.91)
 9. F11=0 138 ==> F6=0 125    conf:(0.91)
10. F18=0 158 ==> OVERALL_DIAGNOSIS=1 143    conf:(0.91)
```

Figure1: Rules generated by Apriori from original dataset



weka.gui.GenericObjectEditor

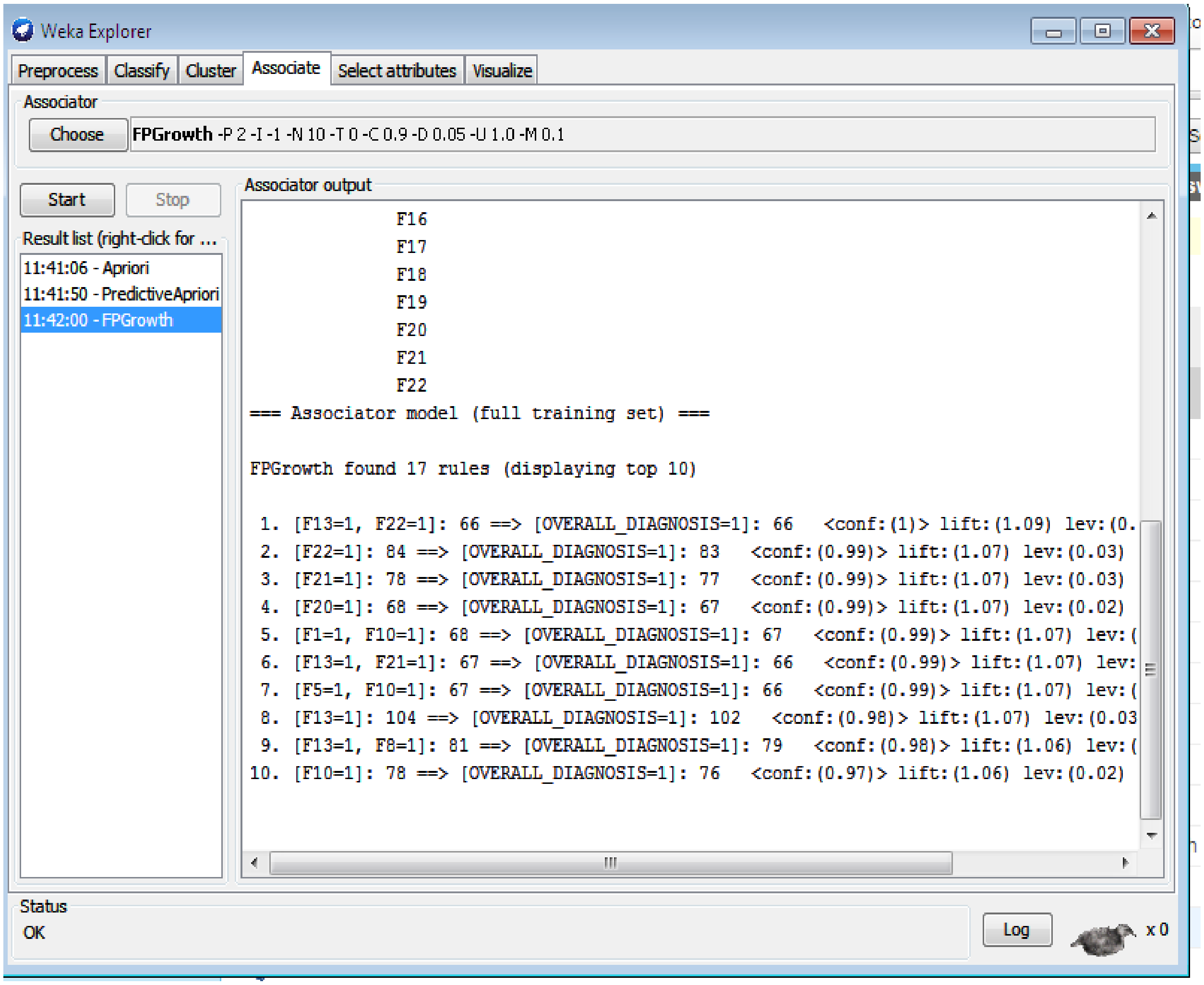
weka.associations.Apriori

About

Class implementing an Apriori-type algorithm.

```
                car   False
         classIndex   -1
              delta   0.05
lowerBoundMinSupport   0.1
         metricType   Confidence
          minMetric   0.9
           numRules   10
     outputItemSets   False
removeAllMissingCols   False
   significanceLevel   -1.0
upperBoundMinSupport   1.0
            verbose   False
```

Figure 2:Properties of Aprori type algorithm FPGrowth:



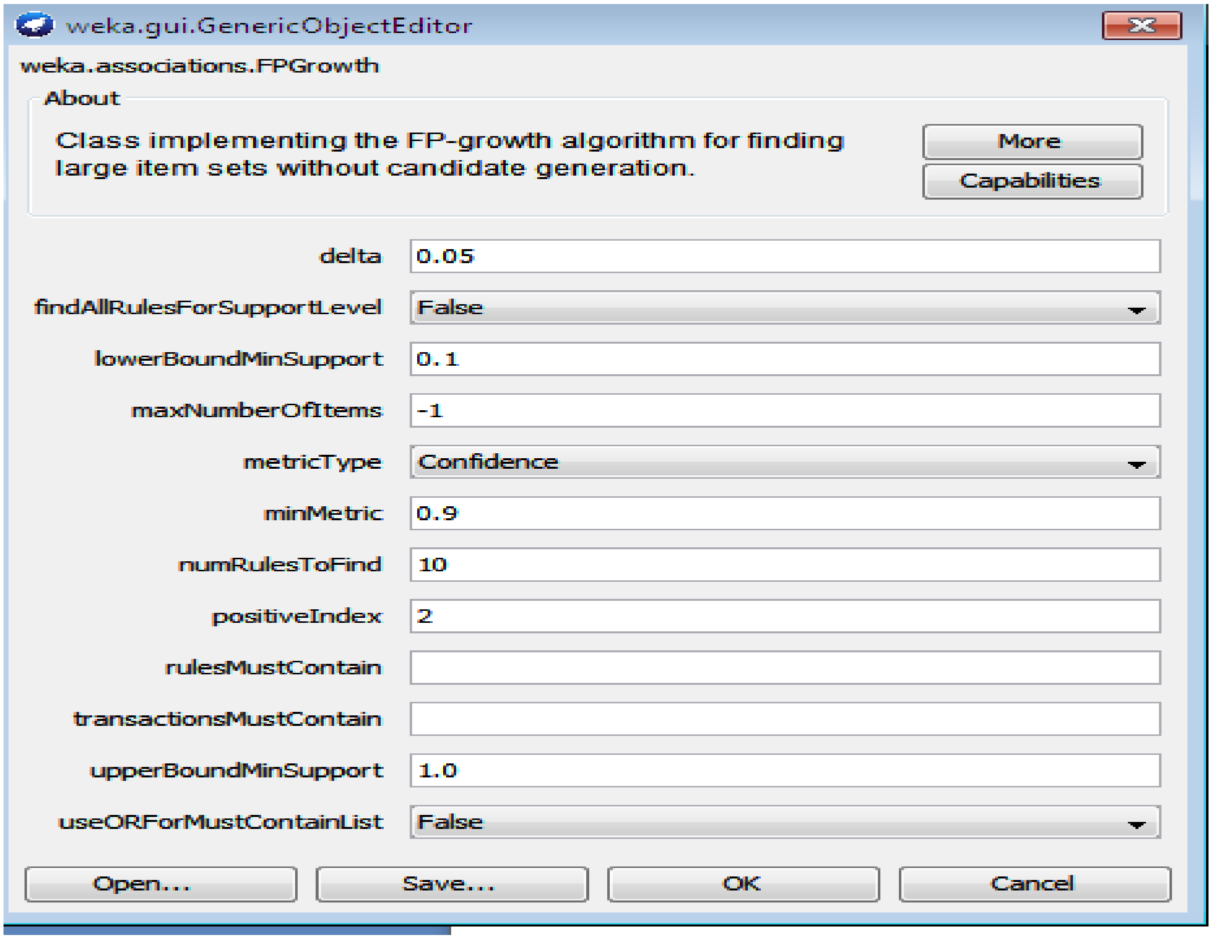
FPGrowth -P 2 -I -1 -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1

11:41:06 - Apriori
11:41:50 - PredictiveApriori
11:42:00 - FPGrowth

```
              F16
              F17
              F18
              F19
              F20
              F21
              F22
=== Associator model (full training set) ===

FPGrowth found 17 rules (displaying top 10)

 1. [F13=1, F22=1]: 66 ==> [OVERALL_DIAGNOSIS=1]: 66   <conf:(1)> lift:(1.09) lev:(0.
 2. [F22=1]: 84 ==> [OVERALL_DIAGNOSIS=1]: 83   <conf:(0.99)> lift:(1.07) lev:(0.03)
 3. [F21=1]: 78 ==> [OVERALL_DIAGNOSIS=1]: 77   <conf:(0.99)> lift:(1.07) lev:(0.03)
 4. [F20=1]: 68 ==> [OVERALL_DIAGNOSIS=1]: 67   <conf:(0.99)> lift:(1.07) lev:(0.02)
 5. [F1=1, F10=1]: 68 ==> [OVERALL_DIAGNOSIS=1]: 67   <conf:(0.99)> lift:(1.07) lev:(
 6. [F13=1, F21=1]: 67 ==> [OVERALL_DIAGNOSIS=1]: 66   <conf:(0.99)> lift:(1.07) lev:
 7. [F5=1, F10=1]: 67 ==> [OVERALL_DIAGNOSIS=1]: 66   <conf:(0.99)> lift:(1.07) lev:(
 8. [F13=1]: 104 ==> [OVERALL_DIAGNOSIS=1]: 102   <conf:(0.98)> lift:(1.07) lev:(0.03
 9. [F13=1, F8=1]: 81 ==> [OVERALL_DIAGNOSIS=1]: 79   <conf:(0.98)> lift:(1.06) lev:(
10. [F10=1]: 78 ==> [OVERALL_DIAGNOSIS=1]: 76   <conf:(0.97)> lift:(1.06) lev:(0.02)
```

OK

Figure3: Rules generated by FPGrowth from original dataset



weka.associations.FPGrowth

Class implementing the FP-growth algorithm for finding large item sets without candidate generation.

delta	0.05
findAllRulesForSupportLevel	False
lowerBoundMinSupport	0.1
maxNumberOfItems	-1
metricType	Confidence
minMetric	0.9
numRulesToFind	10
positiveIndex	2
rulesMustContain	
transactionsMustContain	
upperBoundMinSupport	1.0
useORForMustContainList	False

Figure42: Properties of FPGrowth type algorithm

IV. CONCLUSION

In this paper we have compared two association rule algorithms i.e. Apriori algorithm and FPGrowth. We have analyzed the frequent itemsets generation and number of cycle performed over the Apriori algorithm and Filter Associator in the context of association analysis.

REFERENCES

[1]. Klosgen W and Zytkow J M (eds.), Handbook of data mining and knowledge discovery, OUP, Oxford, 2002.

[2]. Provost, F., & Fawcett, T., Robust Classification for Imprecise Environments. Machine Learning, Vol. 42, No.3, pp.203-231, 2001.

[3]. Larose D T, Discovering knowledge in data: an introduction to data mining, John Wiley, New York, 2005.

[4]. Kantardzic M, Data mining: concepts, models, methods, and algorithms, John Wiley, New Jersey, 2003.

[5]. Smyth P, Breaking out of the Black-Box: research challenges in data mining, Paper presented at the Sixth Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD-2001), held on May 20 (2001), Santra Barbara, California, USA.

[6]. Agrawal, R. and Srikant, R. 1994. Fast algorithms for mining association rules. In Proc. 20th Int. Conf. Very Large Data Bases, 487-499.

[7]. Bocca and C. Zaniolo, editors, Proceeedings of the 20th International Conference on Very Large Data Bases (VLDB’94), pages 475–486, Santiago de Chile, Chile, Sept 1994 . Morgan Kaufmann.

[8]. Scheffer T. Finding Association Rules That Trade Support Optimally against Confidence. Unpublished manuscript.

[9]. Scheffer T. Finding Association Rules That Trade Support Optimally against Confidence. In L. De Raedt and A. Siebes, editors, Proceeedings of the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD’01), pages 424–435, Freiburg, Germany, September 2001. Springer-Verlag.

[10]. Tunedit Machine Learning Repository, Available http://tunedit.org/repo/UCI/spect_test.arff.

[11]. J. Han, H. Pei, and Y. Yin. Mining Frequent Patterns without Candidate Generation. In: Proc. Conf. on the Management of Data (SIGMOD’00, Dallas, TX). ACM Press, New York, NY, USA 2000.