Proceedings of International Conference on Advances in Computer Technology and Management (ICACTM)
In Association with Novateur Publications IJRPET-ISSN No: 2454-7875
ISBN No. 978-81-921768-9- 5
February, 23rd and 24th, 2018

A COMPARATIVE STUDY OF ASSOCIATION RULE MINING ALGORITHMS

MRS. KIRAN TIKAR
Assistant Professor, Dr. D. Y. Patil Institute of MCA, Akurdi, Pune, India[1],
tikar.kiran@gmail.com

DR. KAVITA SURYAWANSHI
Chairman Academics, Dr. D. Y. Patil Institute of MCA, Akurdi, Pune, India[2]
kavita1104@yahoo.com

ABSTRACT
Data mining (DM) techniques is the set of algorithms that helps in extracting interesting patterns and previously unknown facts from larger volume of databases. Today's ever changing customer needs, fluctuation business market and large volume of data generated every second has generated the need of managing and analyzing such a large volume of data. Association Rule mining algorithms helps in identifying correlation between two different items purchased by an individual. Apriori Algorithm and FP-Growth Algorithm are the two algorithms for generating Association Rules. This paper aims at analyze the performance of Apriori and FP-Growth based on speed, efficacy and price and will help in understanding which algorithm is better for a particular situation.

KEYWORDS: MBA- Market Basket Analysis, DM - Data mining, AM- Association Rule Mining, FIS - Frequent Item Set, WEKA- Waikato Environment for Knowledge Analysis

1. INTRODUCTION:

Data Mining is the technique of identifying hidden pattern and previously unknown trends from the large volume of historical data generated from various sources like business transaction, internet etc by using various data mining algorithms. Data mining applications are Retail Industry for Marketing Data analysis, Financial Data Analysis, fraud detection, Biomedical Data Analysis etc. Data mining has vast range of applications ranging from medical field to e-commerce and from education to business field.
Data Mining consist of define a problem statement then collecting and analyzing data against the problem under observation, afterwards deducing a model out of it and testing the model for correctness, finally making corrective changes if required and deploying the model. [1]
Association rule mining was proposed by R. Agrawal in 1993. This model was initially used for Market Basket Analysis (MBA) Association Rule Mining technique is used in Market Basket Analysis to identify customers buying habits and identifying interesting correlations association structures and predictable or unpredictable unknown patterns among the objects collection from the transactional data or various sources of data. Association among objects is analyzed based on confidence and support. Support talks about the occurrences of an object whereas confidence communicates how strong a particular combination of objects group is.
Weka is a tool used for data amnaysis It consist of various machine learning algorithms for data mining. It is written in Java and runs on any platform.
This paper paper emphases on comparative analysis of two Association Rule mining Algorithms i.e. Apriori and FP-growth using WEKA tool. Apriori Algorithm is an iterative process and generates Candidate Frequent Itemsets at every stage, whereas FP-growth generates an n-array Tree for the same.

2. ASSOCIATION RULE MINING

An Association rule can be described as $X \rightarrow Y$ and X and Y are two disjoint set of items. Strength of an Association rule is measured in terms of support and confidence
Association Rule Mining is performed as follows
1. Identify all the frequent item sets having frequency count equal or above minimum Support count.
2. Generate strong Association rules from the frequent Itemset [2]

Support describe frequency of occurrence of an itemset in the given dataset with respect to the total number of records in the dataset.
Confidence is measured through the conditional probability and it describes the reliability on a particular rule deduced.

$$\text{Support, } s(X \longrightarrow Y) = \frac{\sigma(X \cup Y)}{N}$$

$$\text{Confidence, } c(X \longrightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \quad [3]$$

Apriori algorithm is used for frequent item set mining and generating through transactional databases. It starts with identifying the frequent items in the database having set size one and continues till the largest possible set i.e. it performs candidate generation through "bottom up" approach). Afterwards it eliminates the candidates (through pruning step) which are infrequent. If further extensions are not possible then the algorithm terminates.
The FP-Growth Algorithm, proposed by Han, It 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).

3. COMPARATIVE ANALYSIS

WEKA tool is used for performing the comparative analysis between Apriori and FP- Tree Algorithms.
Following screen shots are taken during the phase of analysis.

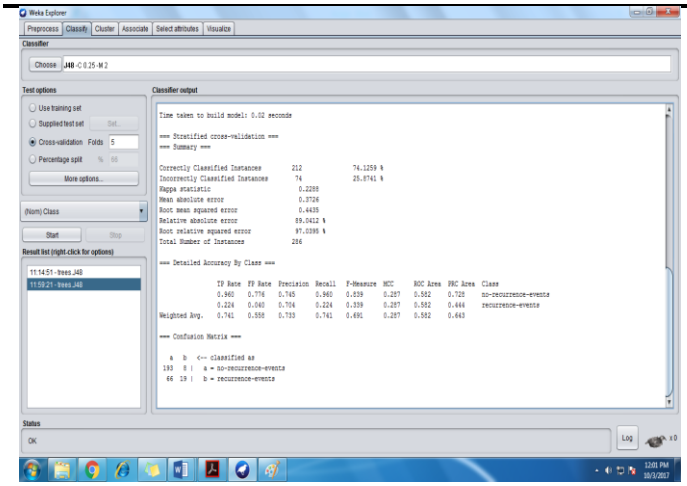

Fig 1 : Result of applying FP-Tree Algorithm on Cancer database in WEKA environment j48 Algorithm

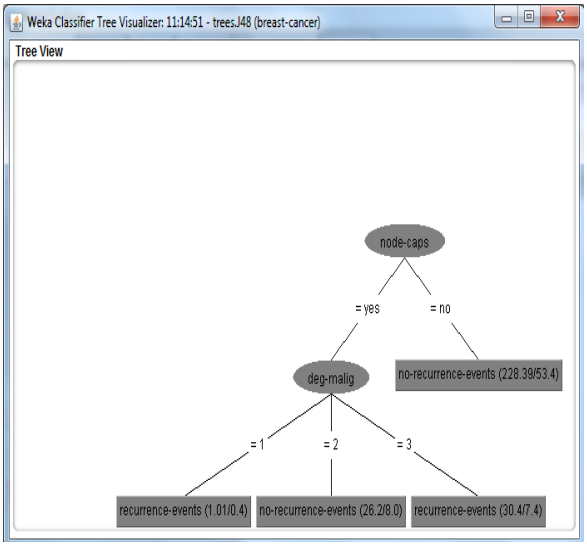

Fig 2 FP-Tree generation as a result of analysis in WEKA environment

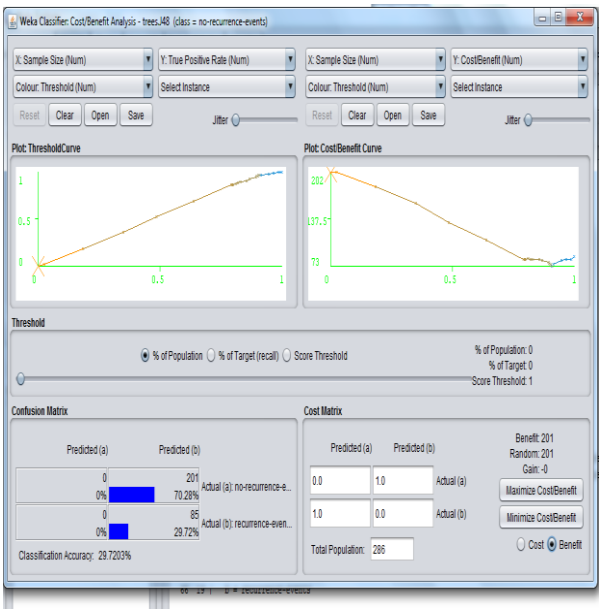

Fig 3: FP-Tree Cost curve generated as a result of analysis in WEKA environment

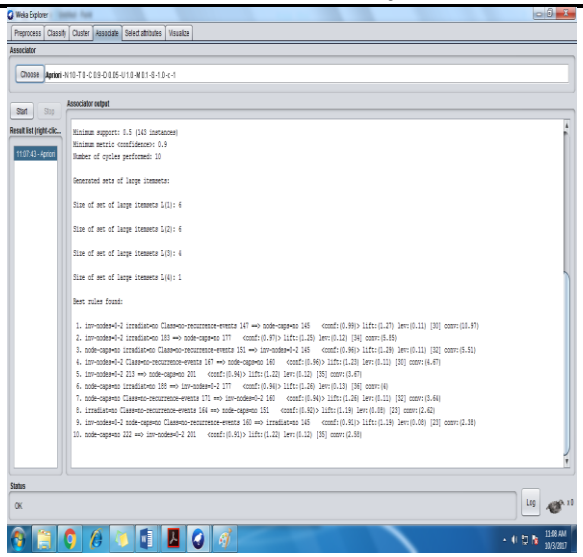

Fig 4: Result of Apriori algorithm analysis in WEKA environment

Difference between FP-Growth and Apriori Algorithm is as follows:

Table1: Comparative Analysis of Apriori and FP-growth Algorithm

S. N.	Apriori Algorithm	FP Growth Algorithm
1	Data items are stored in the form of arrays	Data items are stored in the form of tree.
2	No. of database scans grows with the size of database	Only 2 database scans are required
3	Execution time required is more	Execution time required is less
4	Breadth first search and hash functions are used for searching	Divide and conquer rule is used
5	Joining and Pruning is performed at every stage of execution	An n-array tree is constructed
6	As candidates are generated at each step hence more memory is required	No candidate set generation so less memory is required
7	Mostly used with small size database	Mostly used with medium or large databases
8	Candidate generation is the bottleneck in the performance of Apriori	Better performance as compared to apriori as candidate generation is not required

The following tables present the analysis test results of Apriori and FP-growth for different number of instances and Confidence.

Table 2: Execution time required for different no. of instances

No. of instance	Execution Time(in secs)	
	Apriori	FP-growth
4627	Min Support: 0.15 Confidence: 0.9 No. of Cycles: 17 Time to build: 1.23 sec	No. of leaves : 1 Size of tree: 1 Root relative squared error: 100 % Time to build: 0.06 sec
286	Min Support: 0.5 Confidence: 0.9 No. of Cycles: 10 Time to build: 0.67 sec	No. of leaves : 4 Size of tree: 6 Root relative squared error: 94.6093 % Time to build: 0.06 sec

From the above table it can be identified that when the number of instances decreased, the execution time for

algorithms is also decreased. FP-growth algorithm requires less time as compared to Apriori. So, the performance of FP-growth is better than Apriori.

4. CONCLUSION

The association rules play a major role in many data mining applications, trying to find interesting patterns in databases. The performance analysis of two algorithm for association rule generation is done on the basis of varying number of instances and confidence level. And the efficiency is assessed based on the time required to generate the association rules. The popularity and efficiency of FP-Growth Algorithm contributes with many studies that propose variations to improve his performance [4] [5][6][7][8][9]

FP-Growth is more scalable because of its linear running time. From the analysis it can be concluded that the FP-growth algorithm is cost-effective and proficient as compared to Apriori and steps must be taken to performance of Apriori.

5. REFERENCE:

1) https://msdn.microsoft.com/en-us/library/ ms17 4949. aspx
2) M.S. Mythili and A.R. Mohamed Shanavas, "Performance Evaluation of Apriori and FP-Growth Algorithms" in International Journal of Computer Applications (0975 – 8887) Volume 79 – No10, October 2013
3) https://www-users.cs.umn.edu/~kumar/dmbook/ch6.pdf
4) Agrawal, R. and Srikant, R. 1994. Fast algorithms for mining association rules. In Proc. 1994 Int. Conf. Very Large Data Bases (VLDB'94), Santiago, Chile, pp. 487–499.
5) F. Bonchi and B. Goethals. FP-Bonsai: the Art of Growing and Pruning Small FP-trees. Proc. 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'04, Sydney, Australia), 155–160. Springer-Verlag, Heidelberg, Germany 2004.
6) Cornelia Gyorödi, Robert Gyorödi, T. Cofeey & S. Holban. Mining association rules using Dynamic FP-trees. in Proceedings of The Irish Signal and Systems Conference, University of Limerick, Limerick, Ireland, 30th June 2nd July 2003, ISBN 0-9542973-1-8, pag. 76-82.
7) F. Bonchi, F. Giannotti, A. Mazzanti, and D. Pedreschi. Exante: Anticipated data reduction in constrained pattern mining. In Proc. of PKDD03.
8) Balázes Rácz. Nonordfp: An FP-Growth Variation without Rebuilding the FP-Tree. 2nd Int'l Workshop on Frequent Itemset Mining Implementations FIMI2004.
9) Grahne O. and Zhu J. Efficiently Using Prefix-trees in Mining Frequent Itemsets, In Proc. of the IEEE ICDM Workshop on Frequent Itemset Mining, 2004.