International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in

Vol.3., Issue.4., 2015
(July-Aug)

RESEARCH ARTICLE

ISSN: 2321-7758

COMPARISON OF DATA MINING TOOLS (WEKA vs TANAGRA) USING APRIORI AND FP-GROWTH ASSOCIATION TECHNIQUES

POOJA PANT
Department of Computer Science and Engineering
Amity University , Noida, Uttar Pradesh, India

POOJA PANT

ABSTRACT
Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. It uses machine learning, statistical and visualization techniques to discovery and present knowledge in a form which is easily comprehensible to humans.
Weka and Tanagra data mining toolkit most commonly used. In this paper Weka and Tanagra data mining tools are compared on the basis of Apriori and FP Growth association techniques .

Keywords- Data mining , data warehouse , WEKA data mining tool , TANAGRA data mining tool, Apriori ,FP-Growth

I. INTRODUCTION

This research has conducted a comparison study between WEKA and Tanagra data mining toolkit on the basis of rules discovered by association rules(apriori and frequent pattern)..

The rest of the paper is organized as follows: Section 2 provides an overview of basic data mining Techniques. Section 3 provides an overview of the Data Mining Tools used in our report of experiments. Section 4 provides a brief introduction of association techniques used for comparision. Section 5 shows experiment results. Section .Finally, we close this paper with a summary and an outlook.

Data Mining tasks

There are various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases.

In the next paragraphs we will Summarized only those data mining techniques which are taught to beginners or novice at colleges and schools for understanding the basic concept of mining and finding hidden relationship among various attribute. We have highlighted the specific characteristics for each data mining technique. Some of them can be used to some extent for similar problems, but there are also notable differences that distinguish one from another. The figure 1 shows various data mining techniques taught at basic level.

• **Association** – aims to identify the most common sets of “objects” that appear together, sometimes the end user can choose which objects he wants to be analyzed or how frequent he wants them to be; the user can also set rules of association to better

control the analysis Among the most common applications of association are the identification of products most frequently sold together (market basket analysis), identifying areas where more than two products are sold together most often or the identification of time periods during which the sales are growing.

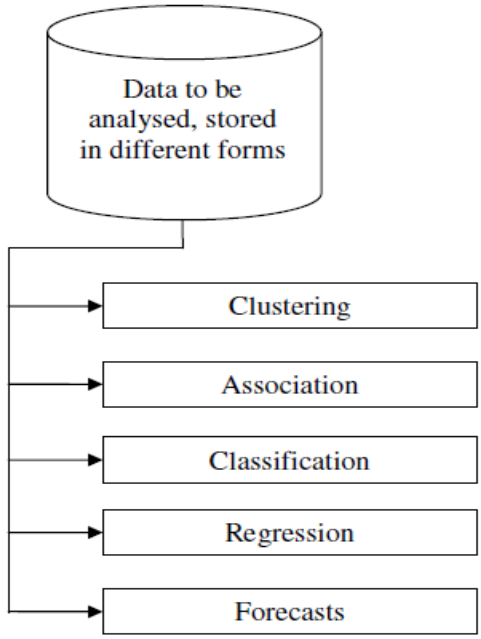


Figure1

• **Classification** – It is one of the most common data mining tasks, responding to issues such as risk analysis; unlike clustering, where categories are not to predetermined characteristics; classification algorithms need to work in a supervised mode, requiring some criteria in order to categorize objects; for this reason usually there must be previously established criteria for classification which can be obtained as a result of the analysis of historical data held; some classification techniques include decision trees, neural networks, etc. As examples can be mentioned the classification of bank customers in terms of how they pay debts, insurance risk analysis in order to determine the insurance premium for various specific cases or establishing property taxes based on certain criteria (value, area, size).

• **Clustering** – helps to find categories of "objects" based on their attributes; groupsare formed containing "objects" similar in many respects (having similar attributes); a difference between clustering and classification is that clustering algorithms can work in an unsupervised mode, taking as input the attributes of the objects and offering as outputs the groups obtained (clusters); Two examples of clustering are grouping the customers of a company according to their purchase value and grouping available products by their features;

• **Forecast** – aims to make a prediction while taking into account past values usually in the form of series of events conducted over time and unlike regression can take into account other factors such as periodical fluctuations of events; As examples of forecasting can be mentioned the seasonal forecasts in different fields of activity (if the variation
is cyclical the regression is not suitable for analysis) or the forecast of sales of goods for the future based on the data available from the past;

• **Outlier analysis** – this type of analysis is intended to detect a number of "objects" that behave very differently from the rest; as an example it is most frequently used for fraud detection, intrusion detection or finding errors;

• **Regression** – is a data mining task that has its origin in statistics where it was used extensively; it resembles from some points of view with classification the difference being the fact that it works on continuous-valued attributes;
regression can be used for predictions also, when the data analyzed is time related, but not for cyclic predictions (linear or polynomial non periodic predictions); a frequent use of regression is the calculation of intermediate values
by interpolation;

• *Sequence analysis* – it is used to identify patterns in a series of discrete values; the main difference between sequence analysis and association is that the first is searching for the order of events and the transitions between different states while the second only searches for correlations between supposed independent objects; an example of sequence analysis can be the identification of models followed by a company's sales over time.

Data Mining Tools

WEKA

WEKA[3], formally called Waikato Environment for Knowledge Learning, is a computer program that was developed at the University of Waikato in New Zealand for the purpose of

identifying information from raw data gathered from agricultural domains. WEKA supports many different standard data mining tasks such as data preprocessing, classification, clustering, regression, visualization and feature selection. The basic premise of the application is to utilize a computer application that can be trained to perform machine learning capabilities and derive useful information in the form of trends and patterns.

WEKA is an open source application that is freely available under the GNU general public license agreement. Originally written in C the WEKA application has been completely rewritten in Java and is compatible with almost every computing platform. It is user friendly with a graphical interface that allows for quick set up and operation. WEKA operates on the predication that the user data is available as a flat file or relation, this means that each data object is described by a fixed number of attributes that usually are of a specific type, normal alpha-numeric or numeric values. The WEKA application allows novice users a tool to identify hidden information from database and file systems with simple to use options and visual interfaces. The following figure 1 presents the WEKA GUI chooser.

Figure 1

TANAGRA

TANAGRA is a free DATA MINING software for academic and research purposes. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area. This project is the successor of SIPINA which implements various supervised learning algorithms, especially an interactive and visual construction of decision trees. TANAGRA is more powerful, it contains some supervised learning but also other paradigms such as clustering, factorial analysis, parametric and nonparametric statistics, association rule, feature selection and construction algorithms. TANAGRA is an "open source project" as every researcher can access to the source code, and add his own algorithms, as far as he agrees and conforms to the software distribution license. The main purpose of Tanagra project is to give researchers and students an easy-to-use data mining software, conforming to the present norms of the software development in this domain (especially in the design of its GUI and the way to use it), and allowing to analyze either real or synthetic data. The second purpose of TANAGRA is to propose to researchers an architecture allowing them to easily add their own data mining methods, to compare their performances. TANAGRA acts more as an experimental platform in order to let them go to the essential of their work, dispensing them to deal with the unpleasant part in the programmation of this kind of tools: the data management. The third and last purpose, in direction of novice developers, consists in diffusing a possible methodology for building this kind of software. They should take advantage of free access to source code, to look how this sort of software is built, the problems to avoid, the main steps of the project, and which tools and code libraries to use for. In this way, Tanagra can be considered as a pedagogical tool for learning programming techniques. The following figure 2 shows the GUI for Tanagra.

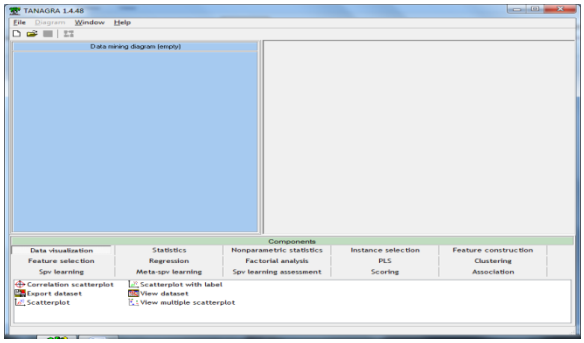


Figure:2

IV Association Techniques

APIORI

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long

as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis.

The pseudo code for the algorithm is given below for a transaction database T, and a support threshold of ϵ. Usual set theoretic notation is employed, though note that T is a multiset. C_k is the candidate set for level k. At each step, the algorithm is assumed to generate the candidate sets from the large item sets of the preceding level, heeding the downward closure lemma. $count[c]$ Accesses a field of the data structure that represents candidate set C, which is initially assumed to be zero. Many details are omitted below, usually the most important part of the implementation is the data structure used for storing the candidate sets, and counting their frequencies.

$\text{Apriori}(T, \epsilon)$
$L_1 \leftarrow \{\text{large } 1-\text{itemsets}\}$
$k \leftarrow 2$
while $L_{k-1} \neq \emptyset$
$C_k \leftarrow \{a \cup \{b\} \mid a \in L_{k-1} \wedge b$
for transactions $t \in T$
$C_t \leftarrow \{c \mid c \in C_k \wedge c \subseteq t$
for candidates $c \in C_t$
$count[c] \leftarrow count[c]$
$L_k \leftarrow \{c \mid c \in C_k \wedge \ count[c]$
$k \leftarrow k+1$
return $\bigcup_k L_k$

FP GROWTH

FP stands for frequent pattern.

In the first pass, the algorithm counts occurrence of items (attribute-value pairs) in the dataset, and stores them to 'header table'. In the second pass, it builds the FP-tree structure by inserting instances. Items in each instance have to be sorted by descending order of their frequency in the dataset, so that the tree can be processed quickly. Items in each instance that do not meet minimum coverage threshold are discarded. If many instances share most frequent items, FP-tree provides high compression close to tree root.

Recursive processing of this compressed version of main dataset grows large item sets directly, instead of generating candidate items and testing them against the entire database. Growth starts from the bottom of the header table (having longest branches), by finding all instances matching given condition. New tree is created, with counts projected from the original tree corresponding to the set of instances that are conditional on the attribute, with each node getting sum of its children counts. Recursive growth ends when no individual items conditional on the attribute meet minimum support threshold, and processing continues on the remaining header items of the original FP-tree.

Once the recursive process has completed, all large item sets with minimum coverage have been found, and association rule creation begins

V Experiment

In this research experiment Vote database was used and 4 experiments(experiment a, b, c and d) were performed using different value of minimum support and confidence . The number of rules discovered was observed in each experiment.

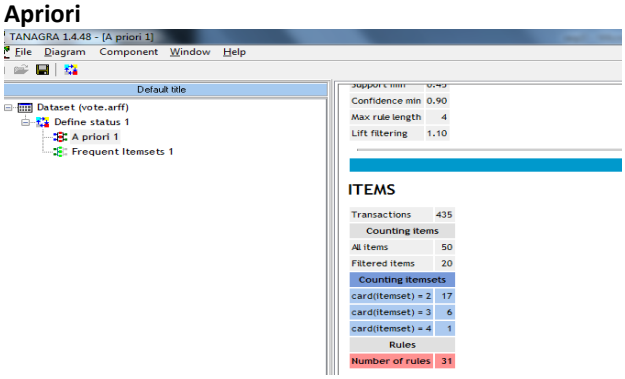
Experiment-a

Minimum support:45%

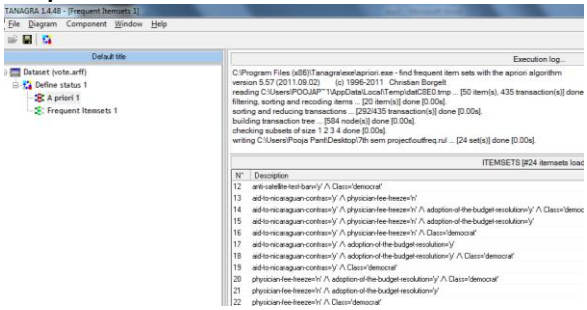
Confidence: 90%

Tanagra

Apriori

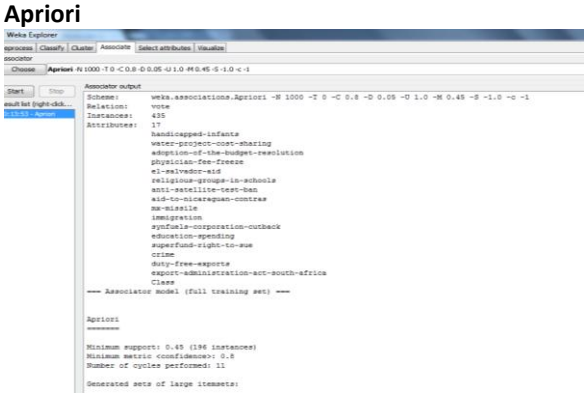


Frequent Itemset

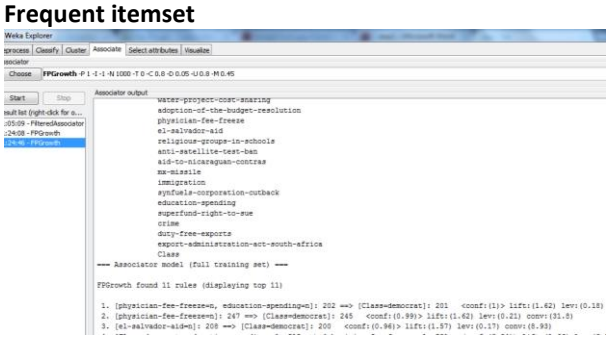


WEKA

Apriori



Frequent itemset



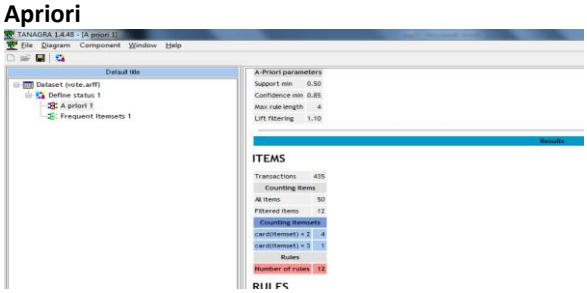
Experiment-b

Minimum support:50%

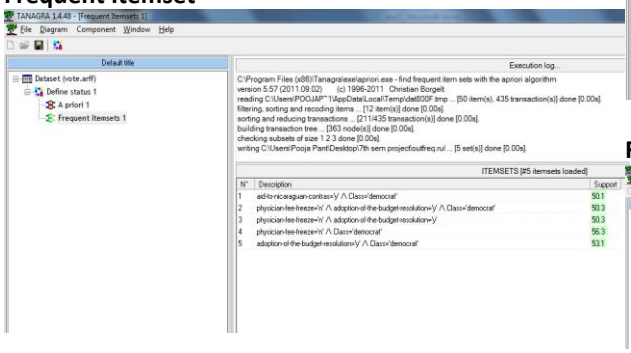
Confidence: 85%

Tanagra

Apriori

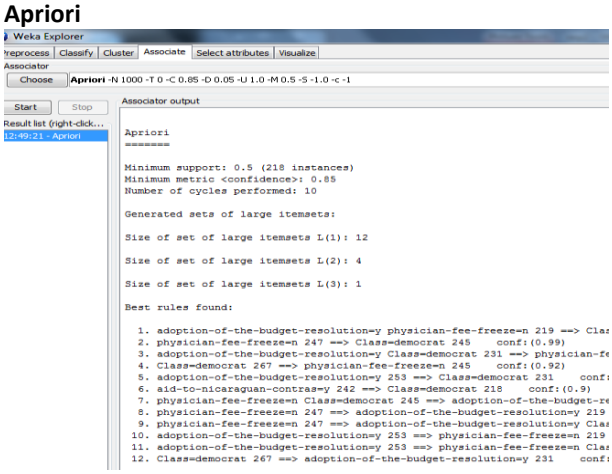


Frequent Itemset

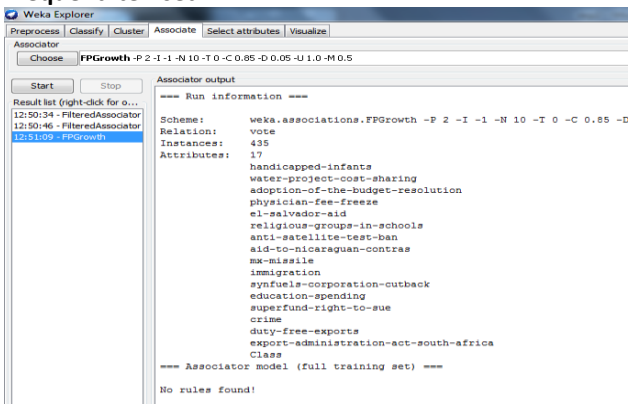


WEKA

Apriori



Frequent itemset



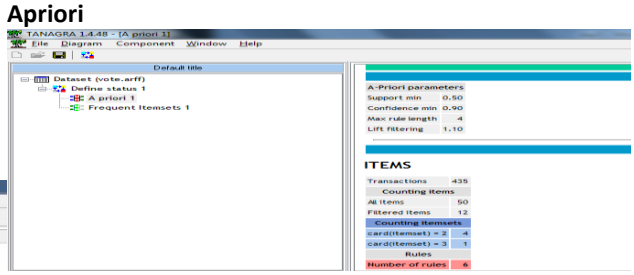
Experiment-c

Minimum support:50%

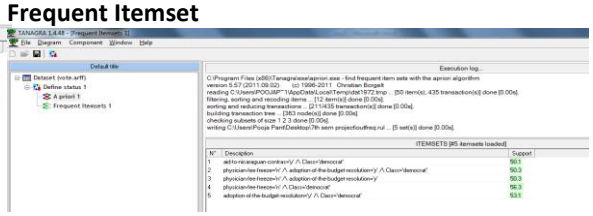
Confidence: 90%

Tanagra

Apriori

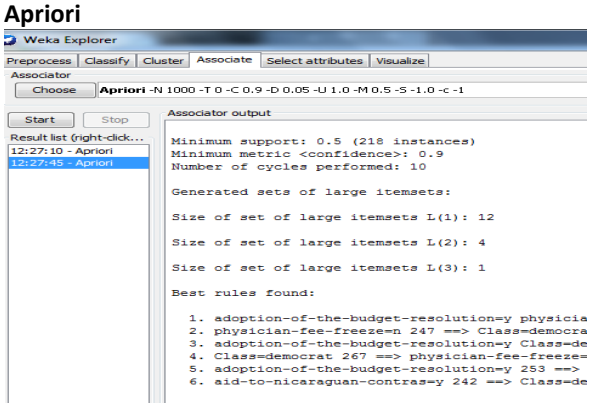


Frequent Itemset

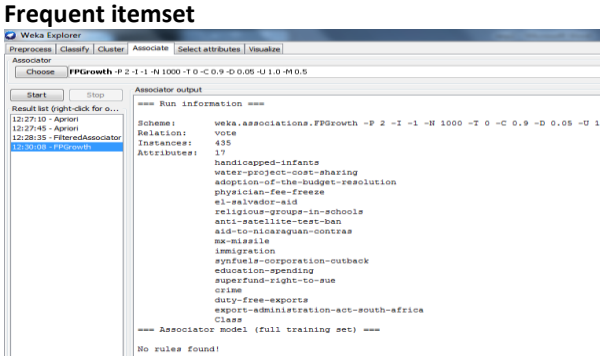


WEKA

Apriori



Frequent itemset



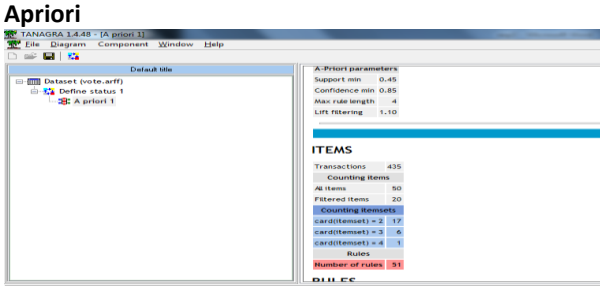
Experiment-d

Minimum support:45%

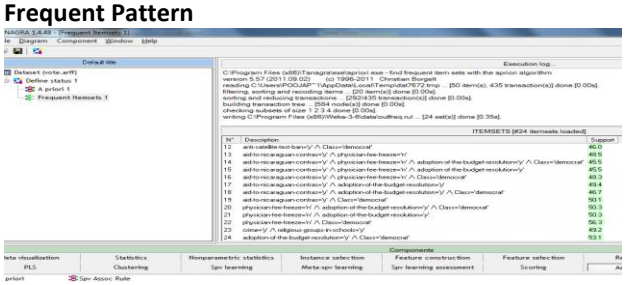
Confidence: 85%

Tanagra

Apriori

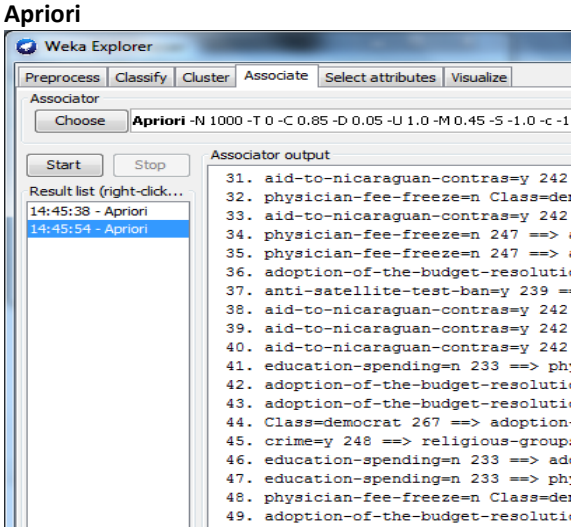


Frequent Pattern

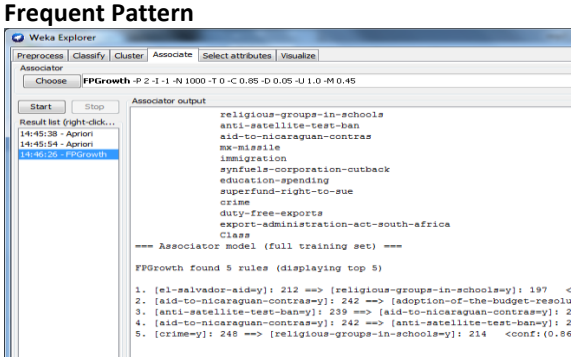


WEKA

Apriori



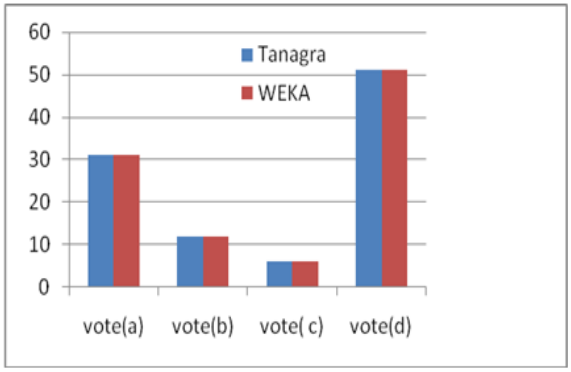
Frequent Pattern



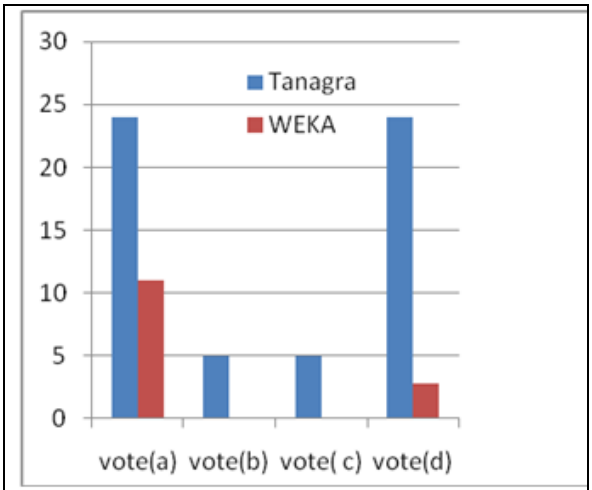
Algorithim	DataTool used	VOTE (45,90)	VOTE (50,85)	VOTE (50,90)	VOTE (45,85)
A Priori	Tanagra	31	12	6	51
	WEKA	31	12	6	51
Frequent Pattern	Tanagra	24	5	5	24
	WEKA	11	X	X	5

Result

Graphical representations of this table is given below



Tanagra(A Priori) vs WEKA(A Priori)



Tanagra(frequent item set) vs WEKA(Frequent Pattern)

VI Conclusion

This research has conducted a comparative study on a dataset between two data mining toolkits (Weka and Tanagra) for the purposes of finding which data mining tool and association technique is better. After analyzing the results of both the tools, we found that the number of rules discovered while using Apriori are the same in both WEKA and Tanagra data tool kit whereas the number of rules discovered using frequent pattern algorithm differ, Tanagra toolkit discovers more rules as compared to the number of rules discovered by weka toolkit. Weka toolkit failed to discover any rules in experiments with high value of confidence and minimum support . This study has concluded that no tool is better than the other if used for Apriori Technique. However,in terms of FP Growth association technique, we concluded that the Tanagra toolkit was a better tool in terms of the discovering more number of rules as compared to Weka.

REFERENCES

[1] Y. Ramamohan, K. Vasantharao, C. Kalyana Chakravarti, A.S.K.Ratnam" A Study of Data Mining Tools in Knowledge Discovery Process" International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-3, July 2012.

[2] Sharon Christa, K. Lakshmi Madhuri, V. Suma" A Comparative Analysis of Data Mining Tools in Agent Based Systems"

[3] Mihai ANDRONIE, Daniel CRISAN "Commercially Available Data Mining Tools used in the Economic Environment" Database Systems Journal vol. I, no. 2/2010

[4] DATA MINING TECHNIQUES AND APPLICATIONS Mrs. Bharati M. Ramageri Abdullah H. Wahbeh, Qasem A. Al-Radaideh, Mohammed N. Al-Kabi, and Emad M. Al-Shawakfa" A Comparison Study between Data Mining Tools over some Classification Methods" (IJACSA) International Journal of Advanced

AUTHORS BIOGRPHY

Ms Pooja Pant is presently pursuing M.Tech in Data Science from Amity University, Noida, Uttar Pradesh. She has done B.Tech degree course in Computer Science and Engineering from Ansal Institute of Technology, Gurgaon in 2015 under the affiliation of the Guru Gobind Singh Indraprastha University. Now she has focused her research interest in various aspects of Data Mining and Warehouse.