International Journal of Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-4, Issue-9)

Research Article | **September 2015**

Comparative Study of Data Mining Algorithms through Weka

[1]Pankaj Singh[*], [2]Sudhakar Singh, [3]Rakhi Garg, [4]Devisha Singh
[1, 2] Dept. of Computer Science, Banaras Hindu University, Uttar Pradesh, India
[3, 4] Dept. of Computer Science, MMV, Banaras Hindu University, Uttar Pradesh, India

Abstract—

With the rapid increase in worldwide information, efficiency of Data Mining Algorithms has been concerned for numerous years. Different data mining algorithms were designed to extract hidden patterns necessary to produce the significant information. These algorithms can be analyzed by using data mining tools. Weka is one of the open source data mining tool available to analyse the performance of data mining algorithm and also helps in its understanding. In this particular paper, we have discussed the results obtained after the execution of various data mining algorithms e.g. association rule mining, clustering and classification on Weka tool. Moreover, we have also done the comparative analysis of different association rule mining, clustering and classification algorithms by using Weka tool in terms of frequent itemsets and clusters generated for a given dataset.

Keywords— Association rule mining (ARM); Clustering; Classification; Weka

I. INTRODUCTION

The aim of data mining is to extract interesting knowledge from the huge database [1]. From the study of extracted patterns, decision-making process can be done very easily. The association rule mining (ARM) algorithms generates frequent itemsets depending on the respective values of Support and Confidence. Classification is the type of data testing with the aim of extracting models describing significant data classes. There are two steps in Data Classification- One are supervised learning where Training data are analysed by classification Algorithm and another is Classification where Test data are used to estimate the exactness of classification rule. Clustering is the method of categorizing the data into clusters in such a way that objects inside a cluster have very high connection in comparison to one another but are very much different to objects in other cluster. We have used Weka tool on several algorithms related to ARM, Clustering and Classification and have also compared them on the basis of their characteristics.

This paper is mainly divided into 4 sections including this. In section 2; we define data mining and its tasks and various methods. In section 3, we discuss the properties of Weka tool and have done the analysis of the results obtained and finally conclude in section 4.

II. DATA MINING AND VARIOUS METHODS

Data mining is a non-trivial removal of implicit, formerly unknown and potentially positive information from data. Data mining is the exploration and analysis by automatic or semi-automatic means of huge extent of data in order to find out meaningful patterns. The origins of data mining are from machine learning/Artificial Intelligence, pattern recognition, statistics, and database systems. There are some Traditional Techniques which may be unsuitable due to enormity of data, high dimensionality of data and heterogeneous nature of data [2, 4, 8, 9].

A. Data Mining Tasks

Prediction and Description Methods are the two methods for data mining tasks. In Prediction Methods, we use some variables to predict unknown or future values of other variables while in Description Methods we find human-interpretable patterns that describe the data [8]. Classification is Predictive while Clustering, Association Rule Discovery are Descriptive [8, 9].

B. Methods of Data Mining

Different data mining algorithms are designed to find hidden patterns necessary to produce the required information. In this paper we discuss only association rule mining, classification and clustering.

1) Classification: Classification is defined as a collection of records also called training set where each record contains a set of attributes and one of the attributes is the class. A form for class feature as a task of the values of new attributes is obtained. The goal of Classification is to generate previously unseen records that should be assigned to a class accurately [10].

2) Clustering: Clustering is defined as a set of known data points, each having a position of attributes, and an equal measure with them. The clusters are obtained in such a way that Data points in solitary cluster are more similar to each other while in separate clusters are very less similar to each other. Euclidean Distance is applied in case of continuous attributes [11].

3) Association Rule Mining (ARM): ARM is defined as a set of records each of which contains some number of items from a given collection and produce dependency rules which will calculate frequency of an item based on frequency of other items. Association rules are useful for analysing and predicting customer behaviour and are widely used in the area of Marketing, customer relationship management etc. [1, 2].

III. WEKA TOOL AND ANALYSIS OF DIFFERENT ARM, CLUSTERING AND CLASSIFICATION ALGORITHMS

Weka is java based open source data mining tool which has group of data mining algorithms like association rules, decision trees and so on. Weka consists of four windows such as Explorer, Experimenter, Knowledge Flow and Simple CLI. Generally, we use Explorer and Experimenter for data mining. For comparison of multiple algorithms, Experimenter is used but for definite results of data mining, Explorer is used. Explorer starts with a screen of data pre-processing as shown in Fig. 1.

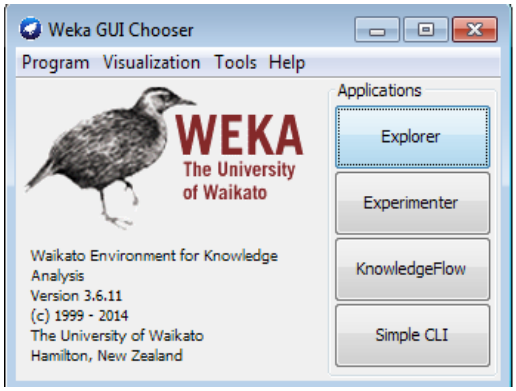

Fig. 1 Outlook of Weka Tool

The basic problem on Weka is opening file because the majority of data sets are in excel format and excel can turn into CSV format. As excel file is semicolon but CSV must have comma, so it needs to be converted into a text file format which is a time taking process [4, 8].

A. ARM Algorithm on Weka Tool

There are different ARM algorithm discovered e.g. Apriori, FP-growth, Eclat, MaxEclat, Clique, Partition etc. We have executed only Apriori and FP-growth algorithms on Weka tool for a given dataset obtained from super market data. The result of Apriori and FP-growth algorithms are shown in Fig. 2, Fig. 3 and Fig 4. We have executed super market dataset by applying Apriori algorithms on Weka tool and obtained the number of frequent itemsets generated for a given value of minimum support and have also applied FP-growth algorithm on Weka tool and obtained number of rules generated by a given value of minimum support and confidence.

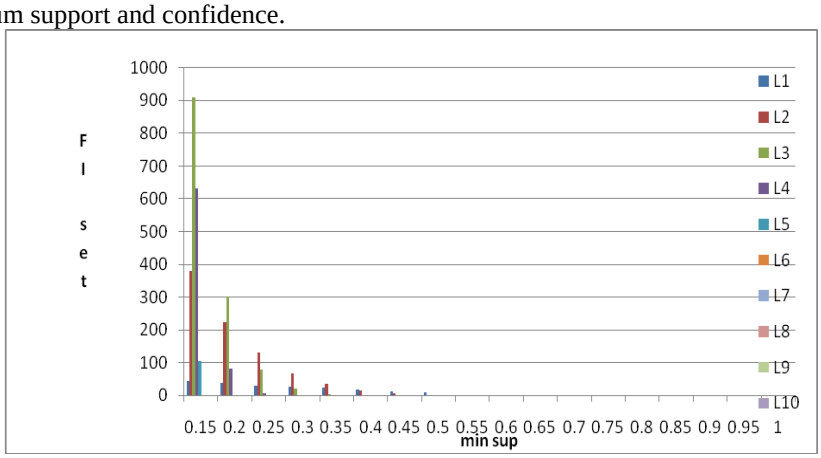

Fig. 2 Apriori Algorithm for different support

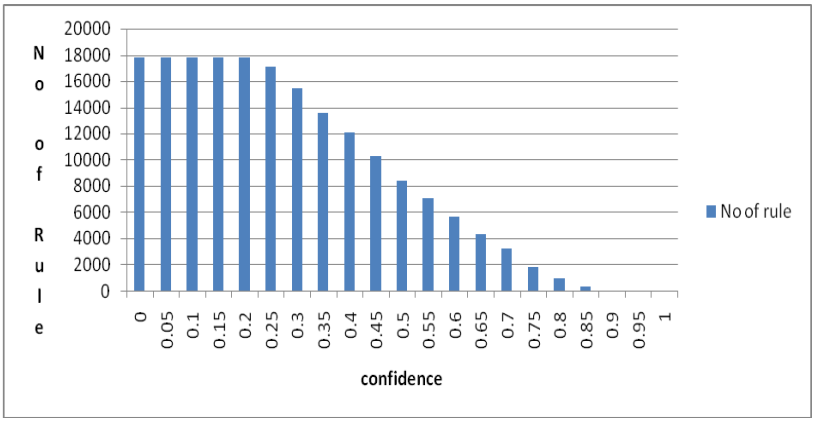

Fig. 3 FP-Growth for different Confidence

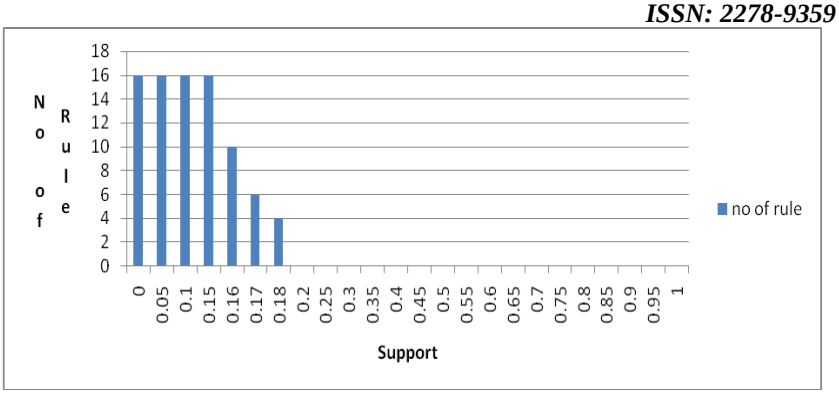

Fig. 4 FP-Growth for different support

In Fig. 2, we have concluded that when we increase the value of minimum support, there is decrease in size of set of large itemsets and frequent itemsets. In Fig. 3 when we increase the value of confidence, numbers of rules generated are decreased. In Fig. 4 when we increase the value of support, we see that at some extent the number of rules generated is same and after that number of rule generated decreases rapidly.

B. Classification on Weka Tool

There are so many algorithms in Classification such as NNge, JRip, M5rules, PART, Ridor, Prism, Conjunctive Rule, Decision table, DTNB, ZeroR and OneR. We have applied Weka tool on Conjunctive Rule, Decision table, DTNB, ZeroR and OneR algorithm on Diabetes data set. The results are shown in figure 5, 6 and 7:-

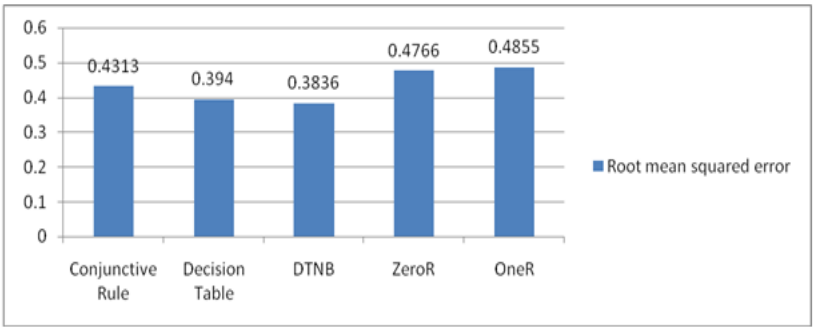

Fig. 5: Root means squared error for different classification algorithms

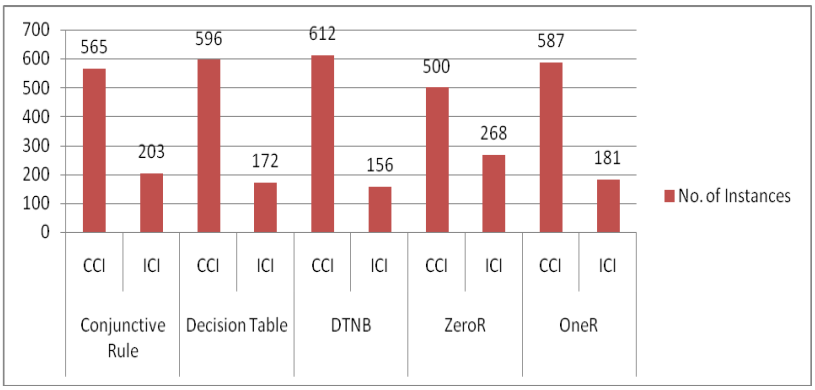

Fig. 6: Number of Instances for different classification algorithms

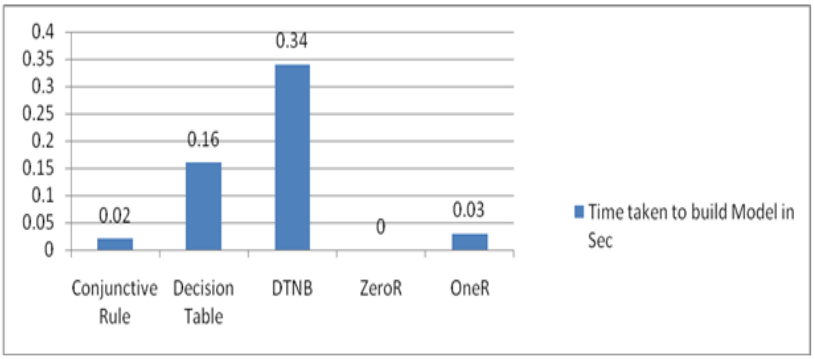

Fig. 7: Time taken to build model in seconds for different classification algorithms

From Figure 5, 6 and 7 we conclude that each decision tree present and achieve a high rate of accuracy. It classify the data into the correctly and the incorrectly instance. We can use these decision tree algorithms in medical, banking, stock market and various areas. In Weka, every data is measured as instances and features in the data are identified as attributes. The imitation results are screened into some sub items for simple study and assessment. Root means squared error in DTNB algorithm are less than other classification algorithm while ZeroR algorithm takes zero second time to build model. The ratio of correctly and incorrect instances in ZeroR algorithm is very less in comparison to other classification algorithms.

C. Clustering on Weka Tool

CLOPE, Cobweb, OPTICS, farthest first, EM, DBSCAN, Simple K-mean and Hierarchical are various Clustering algorithms and we have applied Weka tool on Simple K-mean, Hierarchical, DBSCAN and EM on the diabetes data set. The results are shown below in figure 8, 9 and 10.

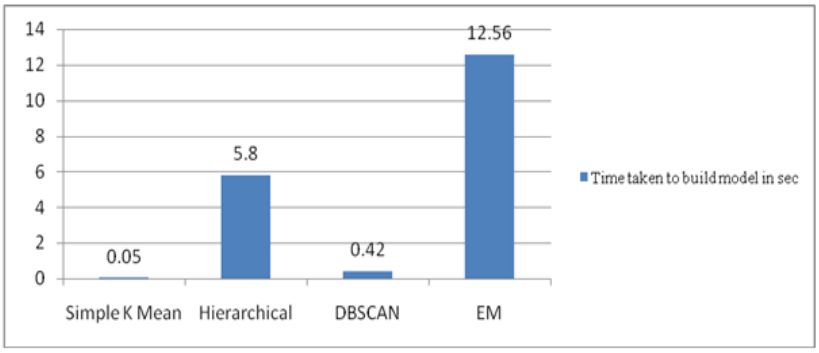

Fig. 8 Time taken to build model in seconds for Clustering

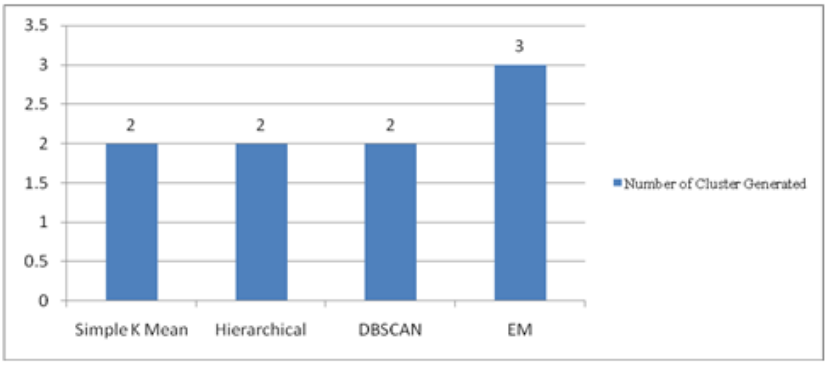

Fig. 9: Number of cluster generated through Clustering

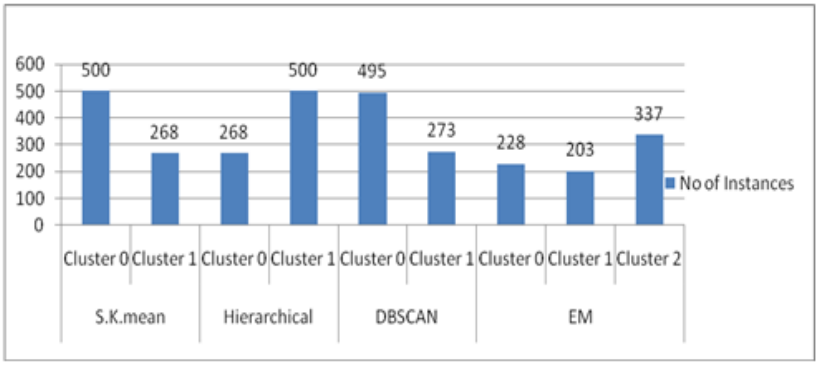

Fig.10: Number of Instances generated by different Clustering Algorithms

We have compared these clustering algorithms on the basis of cluster instances, log likelihood, number of cluster and time needed to build model. The results are shown in Table 1.

TABLE I COMPARISON OF CLUSTERING ALGORITHM

Name of Algorithm	No. of Cluster	Cluster Instances	Time taken to build model in sec.	Log likelihood
Simple K Mean	2	Cluster 0=500 Cluster 1=268	0.05	
Hierarchical	2	Cluster 0=268	5.8	

		Cluster 1=268		
Make Density based cluster	2	Cluster 0=495 Cluster 1=273	0.03	--30.21166
EM	3	Cluster 0=228 Cluster 1=203 Cluster 3=337	12.56	-24.97229

After examine the result we can say that K-Means algorithm generates quality cluster when using huge data set. The performance of Simple K-Mean algorithms is better than Hierarchical Clustering. Density based clustering algorithm is not suitable for data with high variance in density and Hierarchical clustering is more sensitive for noisy data.

IV. CONCLUSIONS

Data mining is the exploration and analysis of huge extent of data in order to find out meaningful patterns. The most popular methods of data mining are classification, clustering and Association rule mining. We executed and evaluated these algorithms on Weka tool. Weka is a java based open source data mining tool, which consist of four windows such as explorer, experimenter, knowledge flow and simple CLI. The advantage of Weka tool is that we need not to know a profound knowledge of programming and implementations. For ARM algorithms we observed that when we increase the value of support number of generated frequent itemset decreases, while in FP-growth number of generated rules is same at some value of support and confidence and decreases after increasing the value of support and confidence. On executing Classification algorithm, we can conclude that each decision tree present and achieve a high rate of accuracy. It classifies the data into correctly and incorrectly instances. Root means squared error in DTNB algorithm are less than other classification algorithm while ZeroR algorithm takes zero second time to build the model. The ratio of correctly and incorrectly instances in ZeroR algorithm is very less in comparison to other classification algorithms. In case of clustering algorithms we found that simple K mean algorithm takes less time than other clustering algorithm. The number of clusters generated by EM algorithm is three and by simple K mean, DBSCAN and Hierarchical algorithms are two. The basic problem with weka is to open a file. Most of the data sets are available in excel format which is not fit for weka tool and hence is to be changed in the required format i.e. arff.

REFERENCES

[1] R. Agrawal and J.C. Shafer, "Parallel mining of association rules," *IEEE Transactions on Knowledge and Data Engineering*, vol.8, no.6, pp. 962-969, Dec. 1996.

[2] M. J. Zaki, "Parallel and distributed association mining: a survey," *Concurrency, IEEE*, vol.7, no.4, pp. 14-25, Oct. - Dec. 1999.

[3] M. J. Zaki, "Parallel and Distributed Data Mining: An Introduction," *Large-Scale Parallel Data Mining*, LNAI 1759, pp. 1-23, Springer-Verlag Berlin Heidelberg 2000.

[4] Xingquan Zhu and Ian Davidson, "Knowledge Discovery and Data Mining: Challenges and Realities", Hershey, New York, 2007.

[5] Meta Group Inc. Data Mining: Trends, Technology, and Implementation Imperatives. Stamford, CT, February 1997.

[6] M. Goebel and L. Grunewald, "A Survey of Knowledge Discovery and Data Mining Tools," *Technical Report, University of Oklahoma*, School of Computer Science, Norman, OK, February 1998.

[7] Waikato ML Group. User Manual Weka: The Waikato Environment for Knowledge Analysis. *Department of Computer Science, University of Waikato* (New Zealand), June 1997.

[8] K. Thearling, Data Mining and Database Marketing WWW Pages. http://www.santafe.edu/~kurt/dmvendors.shtml, 1998.

[9] U. Fayyad, Data Mining and Knowledge Discovery: Making Sense Out of Data, *IEEE Expert*, vol. 11, no. 5, pp. 20-25, October 1996.

[10] Phyu and Nu Thair, "Survey of classification techniques in data mining," in *Proc. International Multi Conference of Engineers and Computer Scientists*, vol. I IMECS 2009, March 18 - 20, 2009, Hong Kong-classification.

[11] G.P. Babu and M.N. Marty, Clustering with evolution strategies Pattern Recognition, vol. 27, no. 2, pp. 321-329, 1994.

