WEKA Powerful Tool in Data Mining

Eshwari Girish Kulkarni
Department of Computer Science & Engineering
Walchand Institute of Technology, Solapur
Maharashtra, India

Raj B. Kulkarni, PhD
Department of Computer Science & Engineering
Walchand Institute of Technology, Solapur
Maharashtra, India

ABSTRACT
The primary task of data mining is to explore the large amount data from different point of view, classify it and finally summarize it. In today's world data mining have progressively become interesting and popular in terms of all application. The data mining requires huge amount of Data sets for extraction of knowledge from it. The main aim of data mining software is to allow user to examine data. This paper is a review paper that introduces the key principle of data pre-processing, introduction of WEKA tool with classification, clustering. "Weka" is a data mining tool. In this paper we are describing the steps of how to use WEKA tool for various technologies & different facility to classify the data through various algorithms.

General Terms
The paper contain some general terms as, general classification of the data, Clustering of Data, Data Pre-processing, Algorithms etc.

Keywords
Data mining, WEKA tool, Data pre-processing, Data set

1. INTRODUCTION
Data mining is a disciplinary sub domain of computer science. Data mining has been defined as the implicit extraction, prior unknown and potentially useful information from historical data databases. It uses machine learning, statistical techniques to discover and represent knowledge in a form, which is simply comprehensive to humans. A number of algorithms have been developed to extract discover knowledge patterns.

Data mining is there search step of the KDD (knowledge discovery in databases process). Data pre-processing, clustering, classification is the popular technologies in data mining. Data mining tools predict behaviors and upcoming trends, helps businesses to make change in knowledge-driven decisions. Data mining tools can solution, business queries that traditionally were too time consuming to resolve.

In this paper the main focus is to detail the ability of the data pre-processing & Weka background.

2. DATA PRE-PROCESSING
2.1 Why to process the Data?
If there is much tangential & redundant information available, noisy & unreliable data and Errors in transmission of data and instruments that collect the faulty data are present then knowledge discovery in Database (KDD) during the training phase is more difficult. Data filtering & preparation can take considerable amount of time. Data pre-processing includes cleaning is the final training set.[1]

2.2 Why Data Pre-processing is & its Methods
Raw data is highly susceptible to noise, lacking attributes values & inconsistency occurred from different data sources. This quality of data infect the DM results. In order to enhance the efficiency & to improve the quality of data &accordingly, the mining results of raw data is pre-processed. Quality decisions must be based on the quality data. **By data processing, standard quality of data can be maintained measured in term of accuracy, completeness, consistency, timeliness, interpretability, believability.**

Duplicates records also need data cleaning.

Data Pre-processing & transformation of the initial dataset. The process of Data Pre-processing are described below:

+**Data Cleaning:-** fill in missing values, resolve inconsistencies & smooth noisy data.

+**Data Integration:-**using multiple databases, or files.

+**Data Transformation:-**aggregation and normalization.

+**Data Reduction:-**reducing the volume but producing similar analytical results.[2]

2.3 Data Pre-Processing Method

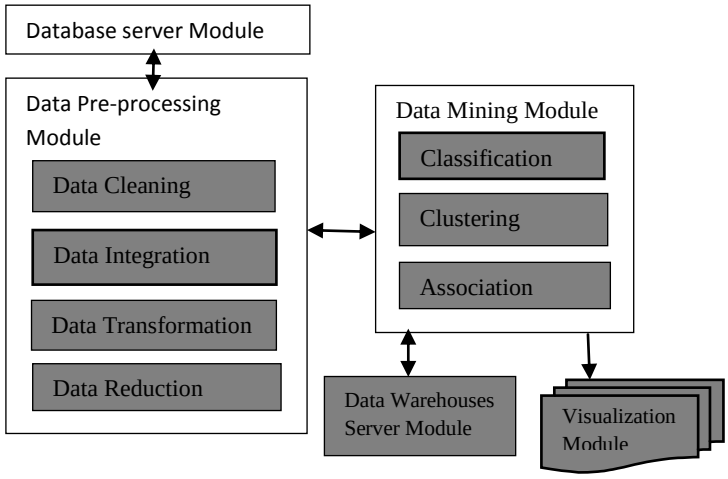

Fig 1: The Steps of Data Pre- Processing Technologies or Method

3. WEKA (WAIKATO ENVIRONMENT FOR KNOWLEDGE ANALYSIS)
Weka is a well-known machine learning software written in Java, developed at Waikato University in New Zealand. 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 Weka workbench contains a collection of visualization tools and algorithms (C4.5 (C5), ID3, K-means, and Apriori) for solving real-world data mining problems and predictive

The workflow of WEKA would be as follows [3]

The workflow of WEKA would be as follows:-
Data →Pre-processing →Data Mining →Knowledge

It has 4 Graphical interfaces [4]

1. **Explorer** is an environment for exploring data
2. **Experimenters** an environment for performing statistical tests between learning schemes.
3. **Knowledge Flow** is Java-Beans for setting up and running machine learning experiments.
4. **Simple CLI** allows direct execution of

Weka commands and also provides simple command-line interfaces.

3.1 Advantages of Weka Include

i. Accessible free under the GNU General Public License.

ii. Portability, as it runs approximately on any modern computing platform.

iii. An exhaustive collection of data pre-processing and modelling techniques.

iv. Due to its graphical user interfaces it is Easy to use.

4. CLASSIFICATION IN WEKA

Purpose: It estimates accuracy of the model.

Classification is the process of finding a set of models that discuss and differentiate data idea and classes, for the purpose of being able to use the model to guess the class whose label is secret.[5]

4.1 Classification is a Two-step Process

i. Build classification model using training data.

ii. The model generated is tested by assigning class labels to data objects data set. The model is represented as decision trees, classification rules and mathematical formulae. It is for classifying future or unknown objects. Accuracy rate is the percentage of independent training set i.e. test set samples that are correctly classified by the model. [5]

We have a data set where each record has attributes A1, A2, A3 upto an, and K. our Goal is to learn a function f: (A1,…,An)K, then use this function to predict k for a given input record (A1,…,An).In this Classification: K is a discrete attribute, called the class label whether Prediction: K is a continuous attribute. Classification Called supervised learning, because true labels (K- values) are known for data provided. Some application involves target marketing, medical diagnosis, and fraud detection. The data is load & Prepare in .arff format.

Classification accuracy: It is the ability to predict categorical class labels. This is the simplest scoring measure. It calculates the proportion of correctly classified instances. [6]

Accuracy = (TP + TN) / (TP+TN+FP+FN)

4.2 Steps Involve in Weka

For the Classification in Weka, we have two categories of classifiers supervised and unsupervised respectively. All the classifiers like tree, rules, lazy, and naïve comes under these categories only. To enhance the accuracy of classifiers Meta classifiers are also present.[6]

Basic four steps for classification in weak

1. Preparing the data
2. Choose classify and apply algorithm
3. Generate trees
4. Analysis the result or output

- **Classification Steps [6]**

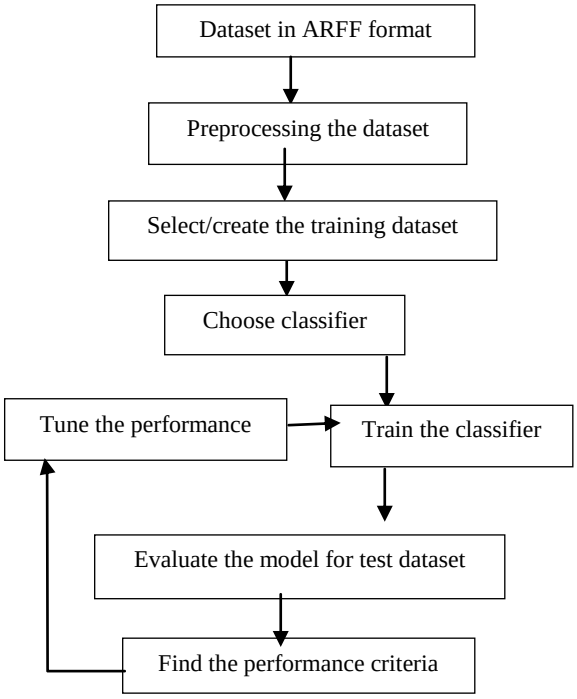

Fig 2: Classification Steps

The classifier model can be improved and prediction strength can be enhanced. It well read (learns), a set of classifiers and combines the predictions of several, classifiers.

Experimental Results

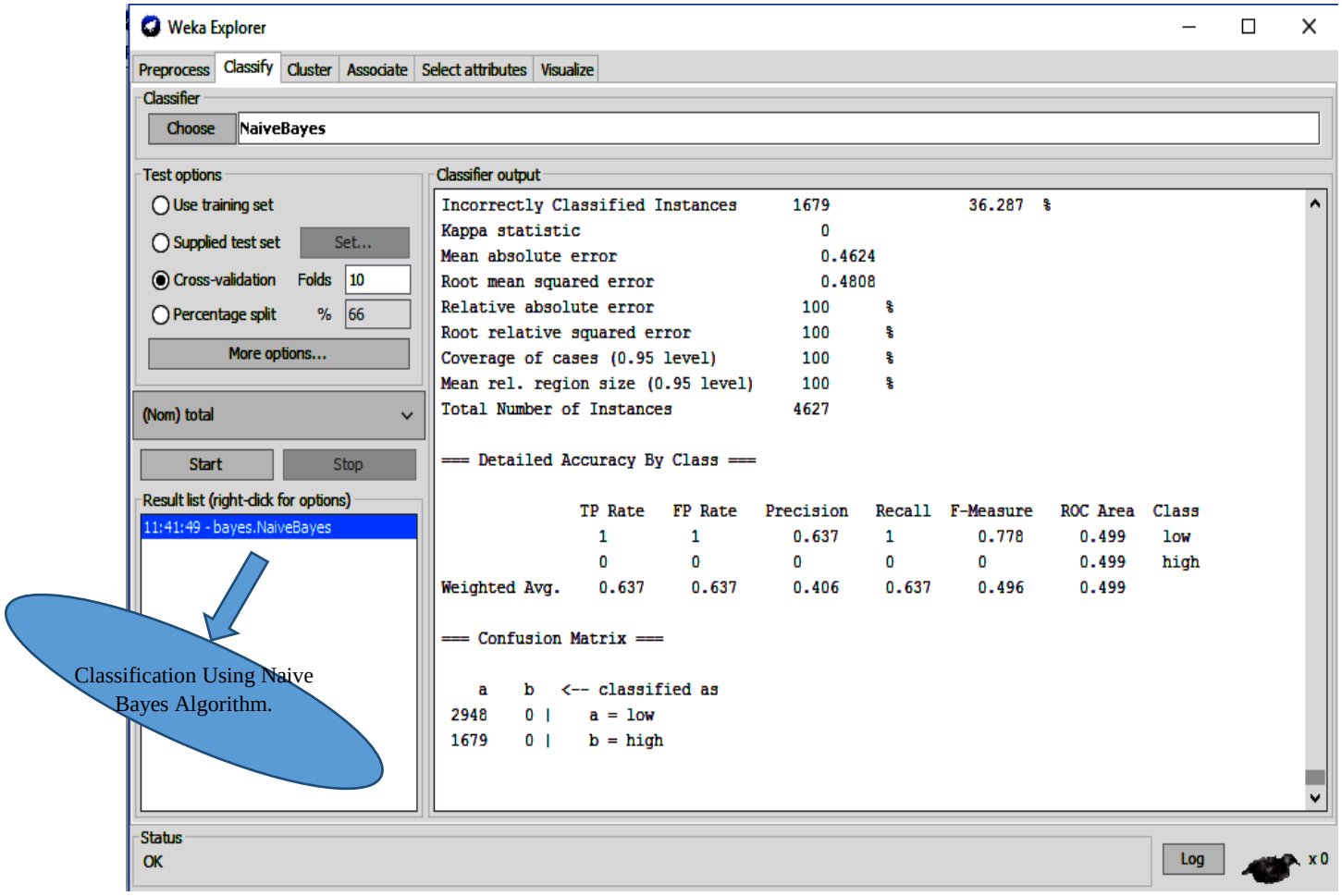

Fig 3: Classification using Naive Bayes Algorithm

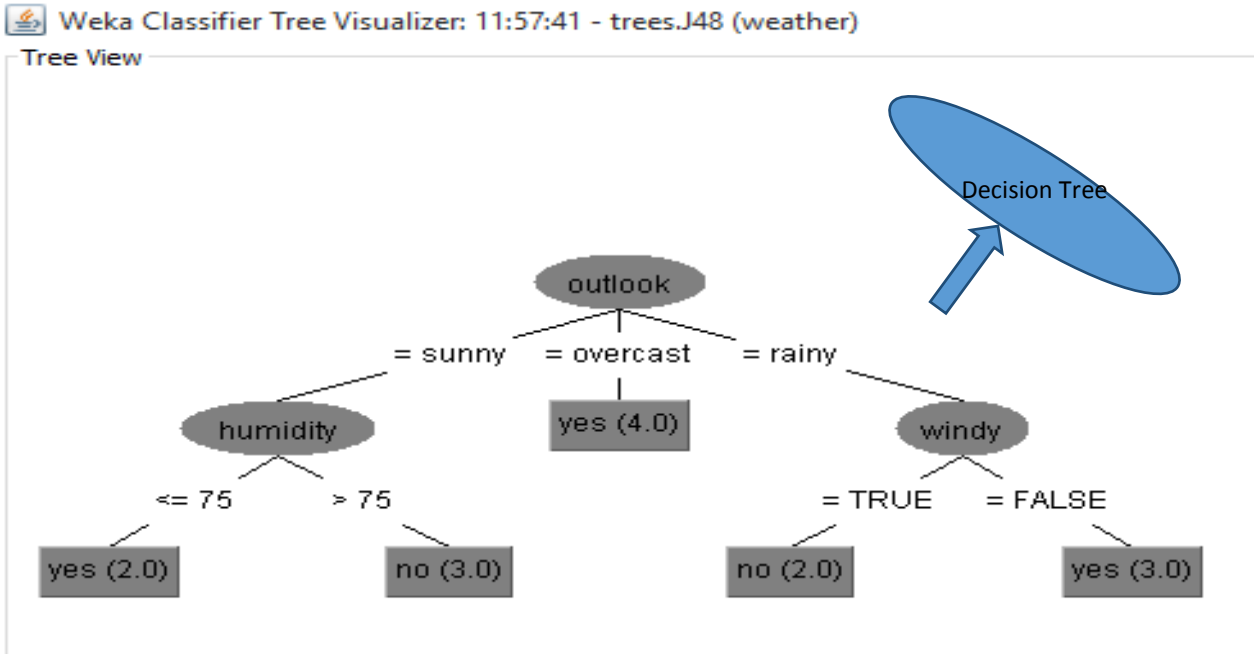

Fig 4: Decision Tree

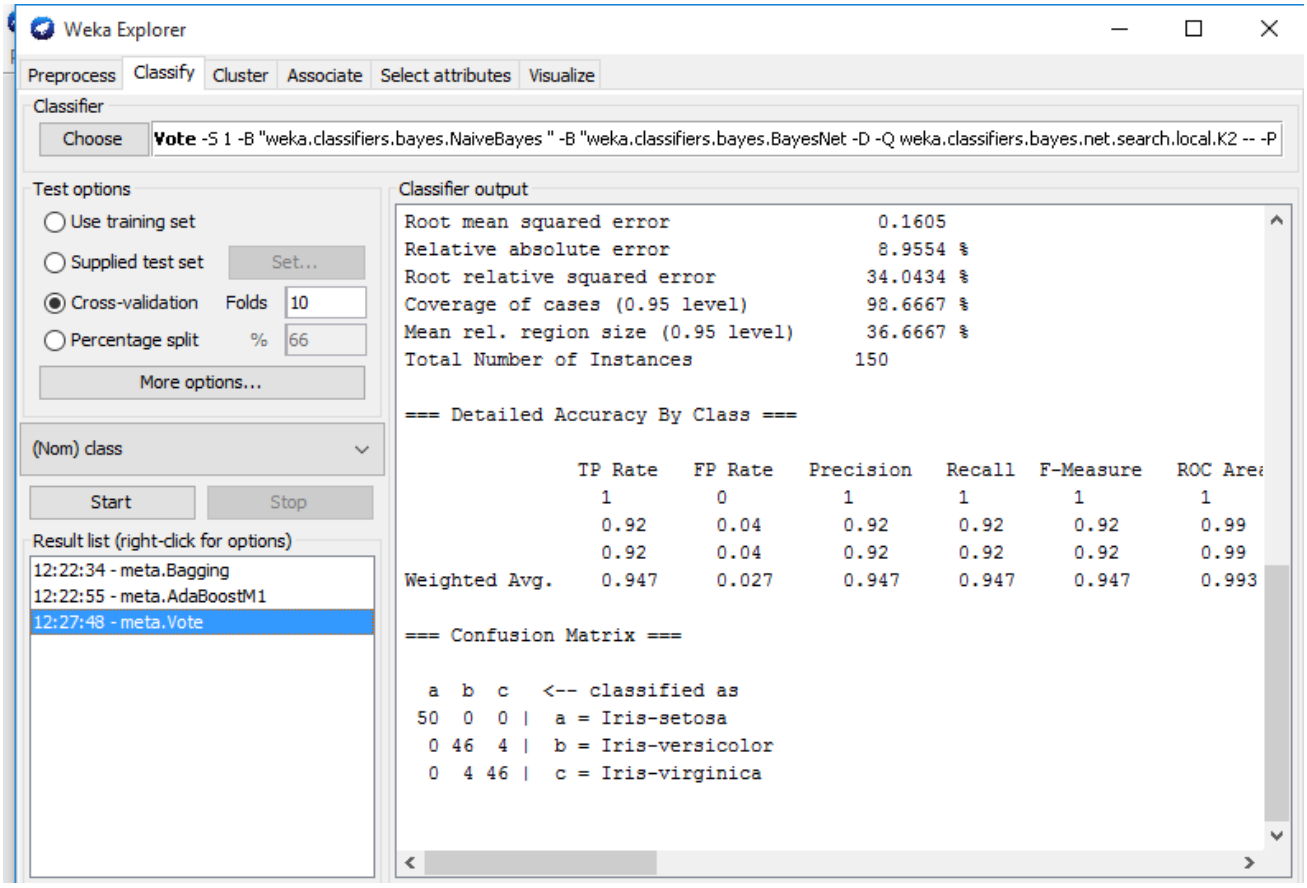

Fig 5: Apply 1st Classification using Naive Bayes Algorithm

Combinining (Merging) Multiple classification Algorithms:- Using two methods of classification approch i.e **Bagging**&**AdboostM1**we can combine multiple classification i.e **Naïve Bayes** and **BayesNet** Algorithmsthatsupport **combination Rules** of "**average probabilities"** For **Reliable** results.

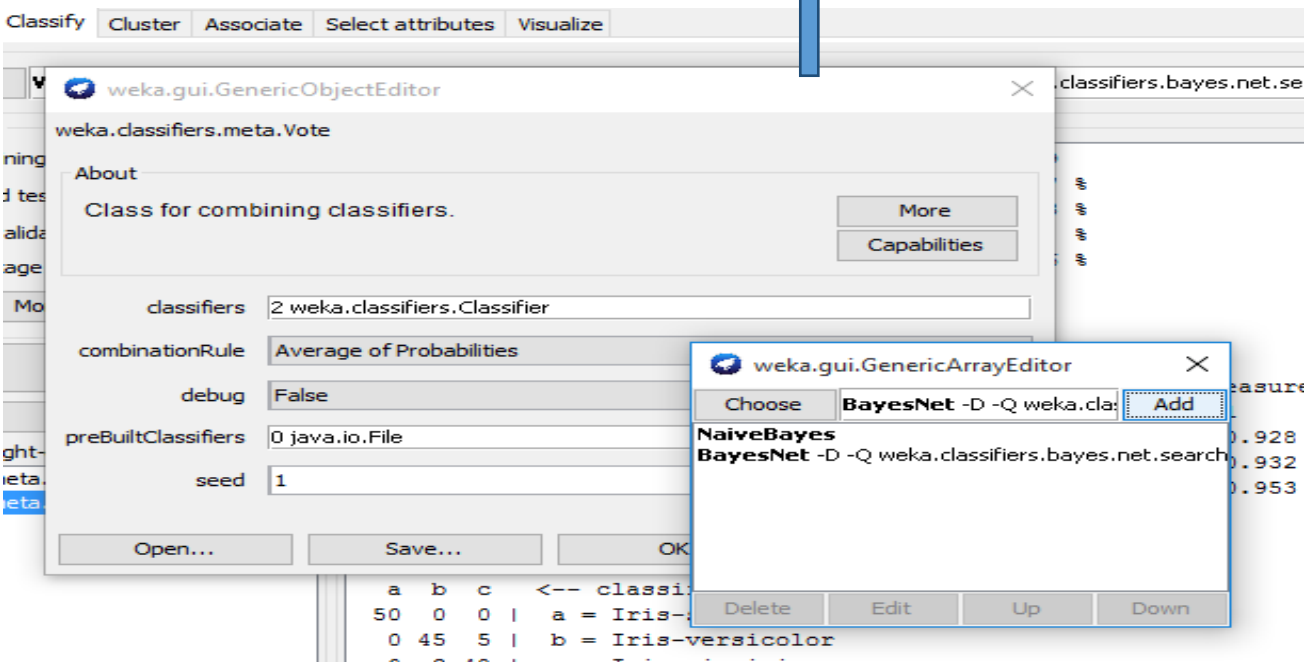

Fig 6: Apply 2nd Classification using Bayes Net Algorithm

5. CLUSTERING IN WEKA

[6]Clustering is used for finding the similar type of objects Different objects are kept in different clusters and group them together in a single cluster. K-means is one of the simplest unsupervised algorithms that solve the clustering problem. Simple Means may be a fruitful, initialize for it valid options are:

N - Identify the number of clusters to generate.

S - Identify random number. It partitions the whole data set in two clusters. Each cluster had shown the lowest and higher value of the data sets**. Advantage of heuristic based method has the time complexity O (nk)**,

Where n = number of objects in the dataset.

K= number of desired clusters.

K-means clustering is a method of cluster analysis in data mining, which aims to partition n observations into k. This results in separate the data space into Verona cells.

There are three major clustering methods and their approach for clustering.

1. K-means Clustering
2. Hierarchical clustering
   i. Agglomerative (bottom up)
   ii. Divisive (top down)
3. Density based clustering
   i. Core points
   ii. 5.3.2 Border points
   iii. 5.3.3 Noise points

Clustering Steps [6]

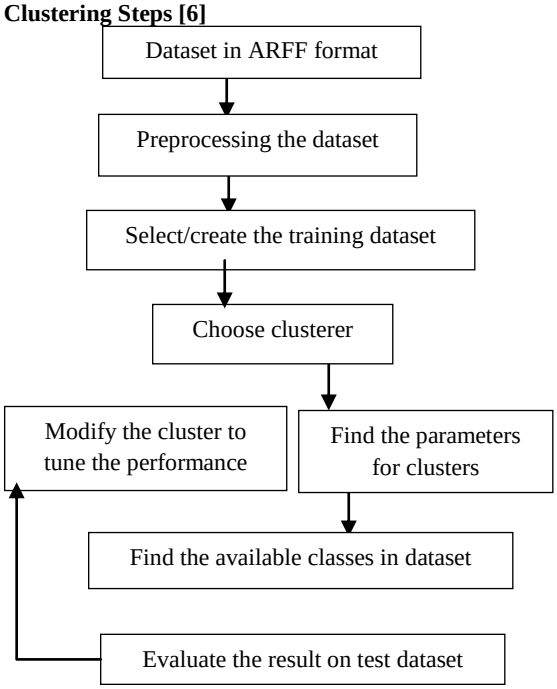

Fig 7: Clustering Steps

In Experimental Results fig (6) and (7) represents the two types of clustering mainly

a) Fig 8: Simple K-Means Clustering
b) Fig 9: Hierarchical Clustering

Experimental Results

a. Clustering using K-Means Algorithm

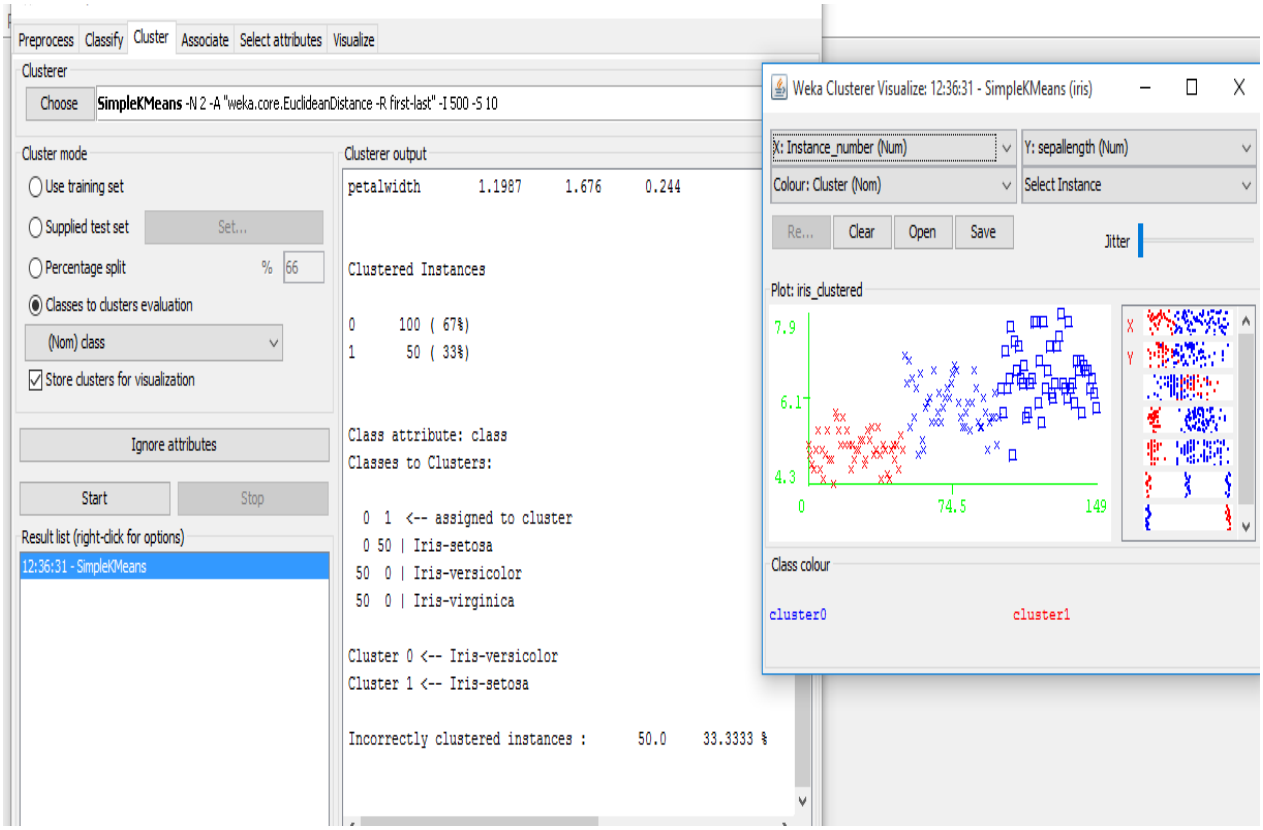

Fig 8: K-Means Clustering

b. Hierarchical Clustering

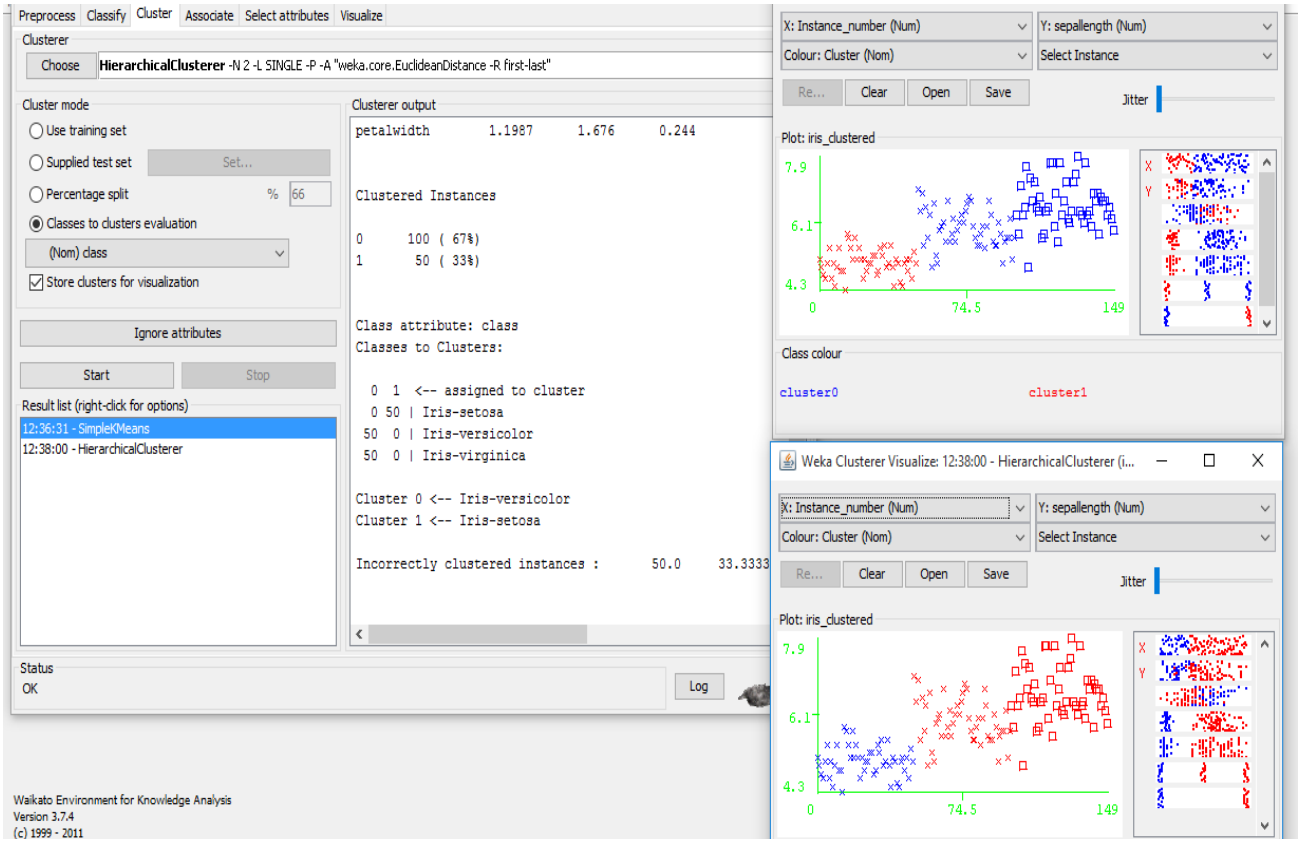

Fig 9: Hierarchical Clustering

6. CONCLUSION

This review paper initially hold brief introduction about the concepts of Data Mining then by moving towards data pre-processing in Weka.

WEKA Powerful tool in Data Mining and Techniques of WEKA such as classification that is used to test and train different learning schemes on the pre-processed data file and clustering used to apply different tools that identify clusters within the data file. This paper also focuses on clustering algorithm such as K-means. The different steps of data mining in Weka

Presented some pictorial representation of experimental results of both accordingly by applying various Algorithms describe in WEKA tool.

7. ACKNOWLEDNMENTS

I am thankful to http://www.cs.waikato.ac.nz/~ml/weka/index.html for providing WEKA tool as open source and some information above is taken from a few related internet sources.

8. REFERENCES

[1] Data Mining Techniques Classification and Prediction—by Han/Kamber/Pei, Tan/Steinbach/Kumar, and Andrew Moore Mirek Riedewald.

[2] Data Mining Algorithms- by Rajanchattamvelli.(Book).

[3] http://www.iasri.res.in/ebook/win_school_aa/notes/Data_Preprocessing.pdf

[4] http://pace.sdsc.edu/sites/pace.sdsc.edu/bootcamp/20140 2/med1ia/PACE_Bootcamp_TS2_WEKA_Intro.pdf

[5] http://www.ittc.ku.edu/~nivisid/WEKA_MANUAL.pdf

[6] http://www.gtbit.org/downloads/dwdmsem6/dwdmsem6l man.pdf

[7] http://www.ijitee.org/attachments/File/v2i6/F084305261 3.pdf.

[8] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1. 1.429.1463&rep=rep1&type=pdf

[9] http://lib.ugent.be/fulltxt/RUG01/000/842/101/RUG01-000842101_2010_0001_AC.pdf

[10] http://www.cs.waikato.ac.nz/~ml/weka/index.html

[11] http://www.ijarcsse.com/docs/papers/Volume_3/9_Septe mber2013/V3I9-0189.pdf