Data Mining with WEKA

Simmi Bagga
[#]*Assistant Professor in Computer*
Sant Hira Dass Kanya MahaVidyalaya, Kala Sanghian
Simmibagga12@gmail.com

***Abstract*— Data Mining is extracting of interesting, non trivial, impact, previously unknown and potential useful information and patterns from the data. WEKA supports many data mining tasks such as data preprocessing, classification, clustering, regression and feature selection. 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. In this paper we elaborate the use of WEKA in Data Mining. We will also explain the WEKA class structure and clustering using K means.**

***Keywords*— Data Mining, WEKA, clustering, K means clustering.**

Introduction

Weka stands for Waikato Environment for Knowledge Learning. It was developed by the University of Waikato, New Zealand. WEKA supports many data mining tasks such as data pre-processing, classification, clustering, regression and feature selection. Data Mining involves a collection of tools and techniques for finding useful patterns relating the fields of very large database.

Data Mining is a powerful tool capable of handling decision making and for forecasting techniques. Data Mining is extracting of interesting, non trivial, impact, previously unknown and potential useful information and patterns from the data. Data Mining task typically involves the analysis of data whose inherent relationship may be obscured by the quality of data. Association of rules as a form of unsupervised learning can be used to extract those relationships so that an analyst can make informed decision based on the available data. Database can be quite large, efficient algorithms of Mining association rules are required to maximize the quality of inferred information and at the same time minimize the computation time. It is the iterative and interactive involves the various steps with the decision made by the user. It uses mathematical analysis to find pattern and trends that exist in data. Clustering techniques are used for combining group that are similar to each other. It is used to find groups or clusters in the data that are similar in some context. Each cluster should be different from other clusters. Clustering is one of the oldest and effective techniques of Data Mining. K-means clustering algorithm is very efficient when we start by finding good initial points.

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. The basic of 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 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.

WEKA is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. It is also well-suited for developing new machine learning schemes. WEKA also works with a unique ARFF file format. Files (data) in the ARFF format are loaded into WEKA for any preprocessing / learning tasks. Regular spread sheets (.CSV files) can also be used to enter data into the system. WEKA has a CSV to ARFF converter built into its system. There are a number of ways to use the data mining algorithms presented in WEKA. The simplest method to do this is by using one of the two graphical user interfaces available in WEKA. These are the Explorer and the Knowledge Flow interfaces. While the Explorer is the perfect platform for users who are not familiar with WEKA to get started, the Knowledge Flow interface is for more experienced users. It provides a better means for visualizing data flow.

A. Main Features of WEKA

- *Weka contains 49 data preprocessing tools.*
- *It has almost 76 classification/regression algorithms.*
- *It contains eight clustering algorithms that performs all the clustering related tasks.*
- *Weka has15 attribute/subset evaluators and almost ten search algorithms for feature selection.*
- *Weka support three graphical user interfaces*

Weka class structure

Every Java program is implemented as a CLASS or collection of classes. In object - oriented programming, a class is a

collection of variables along with some methods that operate on them.

In Weka, the implementation of a particular learning algorithm is encapsulated in a class, and it may depend on other classes for some of its functionality. Some larger algorithms are usually split into more than one class. Rather, the class would include references to other classes that perform parts of the required task. Thus, related classes are grouped together in WEKA in entities known as PACKAGES.

Two packages that are central to WEKA's operation and function are the weka.core package and the weka.classifier package. These packages are not only integral to using machine learning algorithms through WEKA but also for embedding such algorithms in custom Java code.

Clustering in weka

There are several clustering algorithms available in WEKA. These are all part of the WEKA.CLUSTERS package.

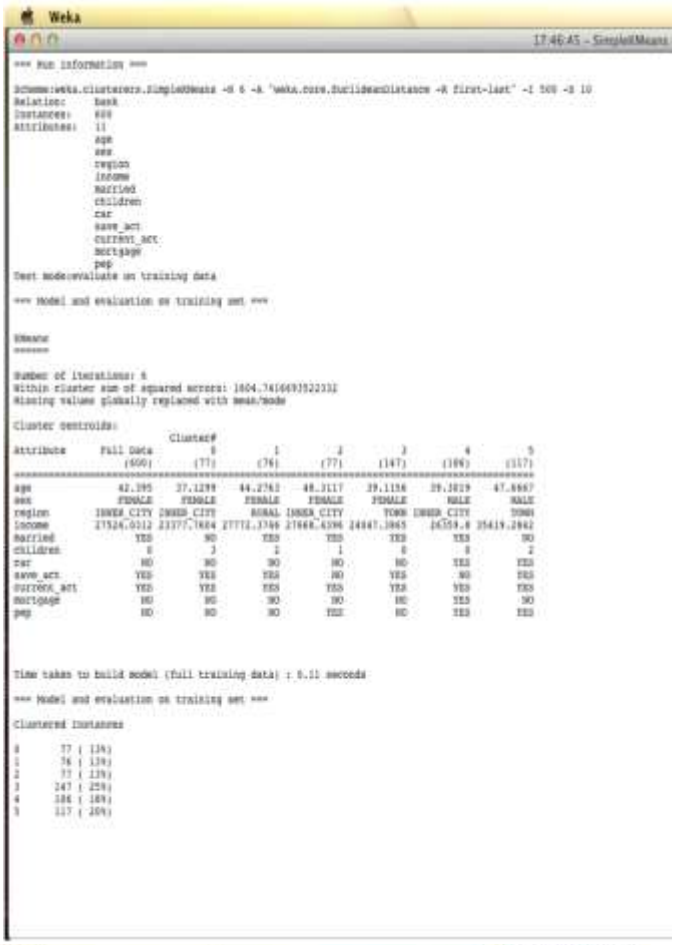
We are interested in K - means clustering, which is defined by the class Simple K Means. Simple K Means clusters data using k-means; the number of clusters is specified by a parameter. The user can choose between the Euclidean and Manhattan distance metrics. In the latter case the algorithm is actually k-medians instead of k-means, and the centroids are based on medians rather than means in order to minimize the within-cluster distance function. (WEKA Manual).

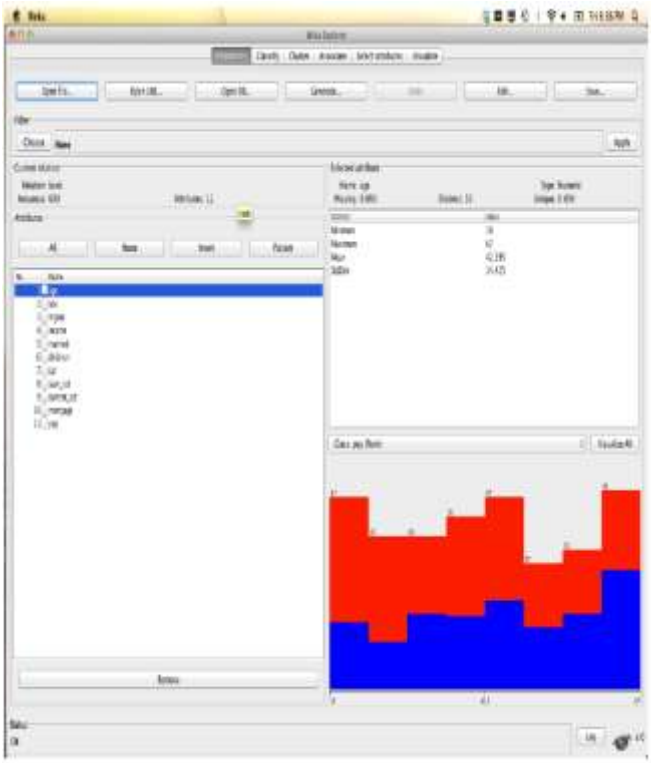
The table below lists the various clustering algorithms available in WEKA.

NAME	FUNCTION
CLOPE	Fast clustering of transactional data
Cobweb	Implements the Cobweb and Classit clustering algorithms
DBScan	Nearest-neighbor-based clustering that automatically determines the number of clusters
EM	Cluster using expectation maximization
FarthestFirst	Cluster using the farthest first traversal algorithm
FilteredClusterer	Runs a clusterer on filtered data
HierarchicalClusterer	Agglomerative hierarchical clustering
MakeDensityBasedCluster	Wrap a clusterer to make it return distribution and density
OPTICS	Extension of DBScan to hierarchical clustering
sIB	Cluster using the sequential information bottleneck algorithm
SimpleKMeans	Cluster using the *k*-means method
XMeans	Extension of *k*-means

USING K-MEANS CLUSTERING IN WEKA

In order to better understand the k - means clustering algorithm, to see how to use this algorithm within the WEKA framework as well as to understand how WEKA handles data, a dataset of bank information was taken and subjected to the k- means clustering algorithm. The value of k was user defined (default value = 2), value of 6 was selected for this experiment**.**





The above screenshot shows that for this data, there are a total of 600 instances and 11 attributes in total. The above snapshot shows a summary of the processing of the k-means algorithm on the loaded data.

WEKA also provides a means of visualizing the data. This provides a powerful means for understanding how the data is being clustered. The above screenshot shows a plot of region vs income. The different clusters can be clearly seen. On closer inspection of this data alone it would appear that possibly a larger value of k would yield a better, more efficient solution.

I. CONCLUSIONS

The WEKA machine learning toolbox is an excellent way for new comers to the field of machine learning to get started. The Explorer and Knowledge flow interfaces provide a relatively simple and intuitive GUI which easy to use and navigate. The core of the WEKA toolbox however lies in the command line interface. Learning this mode provides insight to the class structure of WEKA and plays an important role in embedding WEKA's algorithms in custom Java code. Some of the workflows in both the Explorer and the Command line interface are a little difficult to understand and interpret without prior knowledge of the subject.

REFERENCES

1. Han Jiawei and Kamber, Micheline (2001). "Data Mining: Concepts and Techniques". Morgon Kaufmann. Sanfransico, CA.
2. Singh G.N, Bagga Simmi (2011), Clustering Method for categorical and Numeric Data sets, *Global Journal in computer Science*.
3. Singh G.N, Bagga Simmi (2011), Three Phase Iterative Model of KDD, *International Journal of Information Technology and Knowledge Management*, Volume 4, No.2, pp.695-697.
4. Yao, Y.Y., Zhong, N. and Zhao, Y. (2004), A three-layered conceptual framework of data mining, Proceedings of IEEE ICDM'04 Workshop on Foundations of Data Mining, 205-212.

