A Case Study On Weka

Anushka Singh

15070121109

INTRODUCTION

Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. 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 Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

Weka is free software available under the GNU General Public License.

WEKA 3(Last Update: 2014-09-23),the latest version of WEKA has been optimized to handle BIG Data.

It contains a collection of visualization tools and algorithms for data analysis and predictive modeling, ...together with graphical user interfaces for easy access to this functionality. The original non-Java version of Weka was a TCL/TK front-end to (mostly third-party) modeling algorithms implemented in other programming languages, plus data preprocessing utilities in C, and a Makefile-based system for running machine learning experiments. This original version was primarily designed as a tool for analyzing data from agricultural domains,but the more recent fully Java-based version (Weka 3), for which development started in 1997, is now used in many different application areas, in particular for educational purposes and research.

Advantages of Weka include:

- free availability under the GNU General Public License
- portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform
- a comprehensive collection of data preprocessing and modeling techniques
- ease of use due to its graphical user interfaces

Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported). Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query. It is not capable of multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table that is suitable for processing using Weka. Another important area that is currently not covered by the algorithms included in the Weka distribution is sequence modeling.

User interfaces

Weka's main user interface is the Explorer, but essentially the same functionality can be accessed through the component-based *Knowledge Flow* interface and from the command line. There is also the *Experimenter*, which allows the systematic comparison of the predictive performance of Weka's machine learning algorithms on a collection of datasets.

The *Explorer* interface features several panels providing access to the main components of the workbench:

- The *Preprocess* panel has facilities for importing data from a database, a CSV file, etc., and for preprocessing this data using a so-called *filtering* algorithm. These filters can be used to transform the data (e.g., turning numeric attributes into discrete ones) and make it possible to delete instances and attributes according to specific criteria.
- The *Classify* panel enables the user to apply classification and regression algorithms (indiscriminately called *classifiers* in Weka) to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions, ROC curves, etc., or the model itself (if the model is amenable to visualization like, e.g., a decision tree).
- The *Associate* panel provides access to association rule learners that attempt to identify all important interrelationships between attributes in the data.
- The *Cluster* panel gives access to the clustering techniques in Weka, e.g., the simple k-means algorithm. There is also an implementation of the expectation maximization algorithm for learning a mixture of normal distributions.
- The *Select attributes* panel provides algorithms for identifying the most predictive attributes in a dataset.
- The *Visualize* panel shows a scatter plot matrix, where individual scatter plots can be selected and enlarged, and analyzed further using various selection operators.

Package Manager

In the latest update,Weka 3.7.2,a package manager was added to allow the easier installation of extension packages. This change makes it easier for other Developers to contribute extesions to Weka and to maintain the software, as this modular architecture allows independent updates of the Weka core and individual extensions.This makes addition of new features quick and easy.

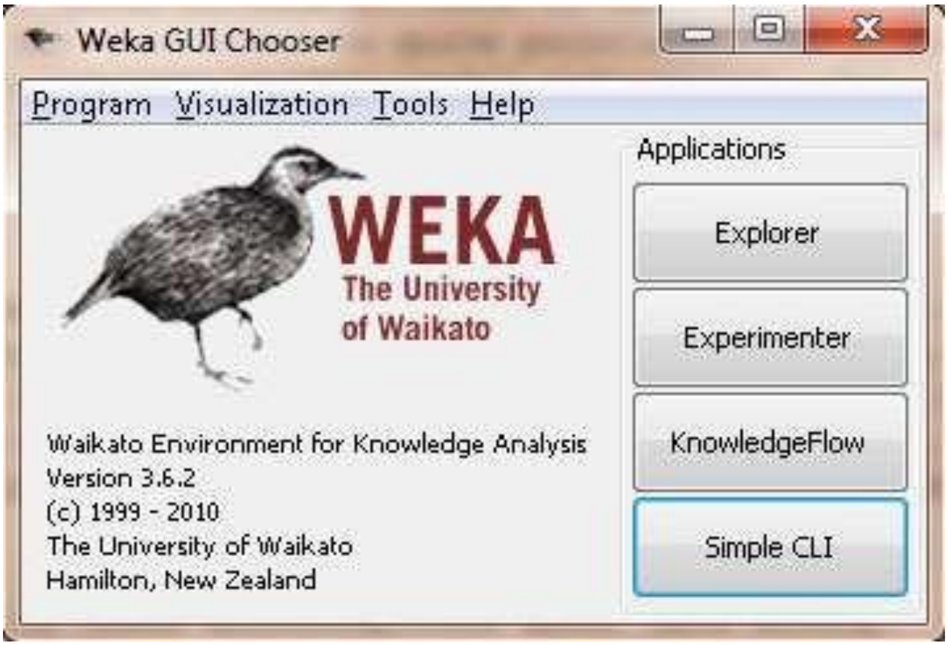
Application

They worked on:

- predicting the internal bruising sustained by different varieties of apple as they make their way through a packing-house on a conveyor belt;
- predicting, in real time, the quality of a mushroom from a photograph in order to provide automatic grading;
- classifying kiwifruit vines into twelve classes, based on visible-NIR spectra, in order to determine which of twelve pre-harvest fruit management treatments has

been applied to the vines;

- Weka has been used extensively in the field of bioinformatics. Published studies include automated protein annotation , probe selection for gene expression arrays , plant genotype discrimination , and classifying gene expression profiles and extracting rules from them.
- Text mining is another major field of application,and the workbench has been used to automatically extract key phrases from text , and for document categorization, and
- word sense disambiguation .
- There are many projects that extend or wrap WEKA in some fashion. There are 46 such projects listed on the Related Projects web page of the WEKA site3.



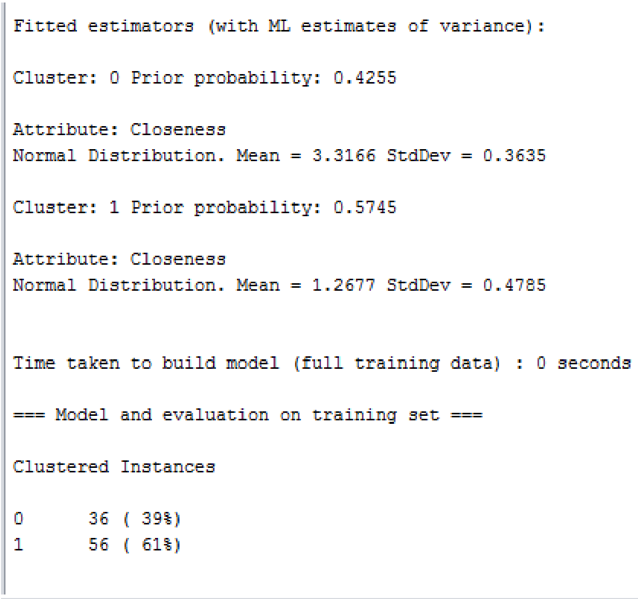
Waikato Environment for Knowledge Analysis
Version 3.6.2
(c) 1999 - 2010
The University of Waikato
Hamilton, New Zealand

Screenshots:

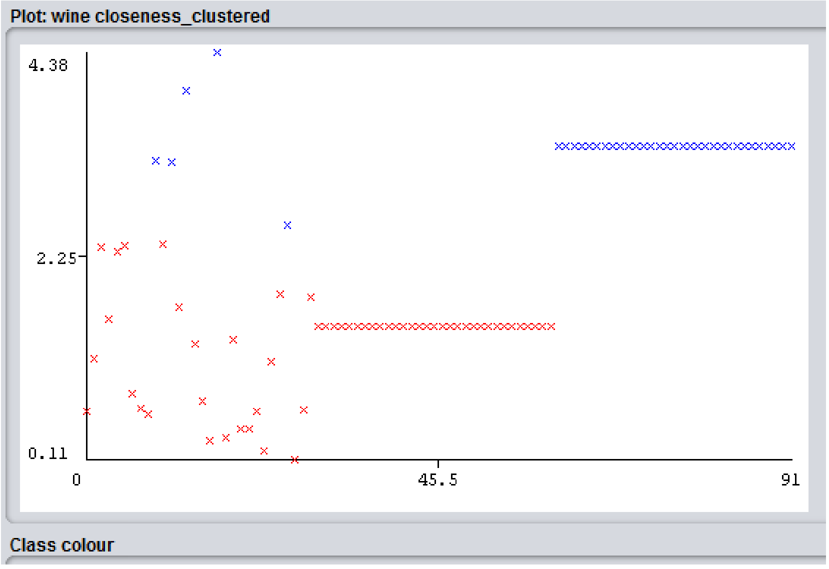
It helped me a lot in my project for the analysis based on cluster formation.

My project is an implementation of CFBA (Closeness Factor Based Algorithm) in a distributed environment, which is an innovation in the field of “advance machine learning” by providing a comprehensive package which forms quality clusters based on closeness between given data series. CFBA is a cluster first algorithm.

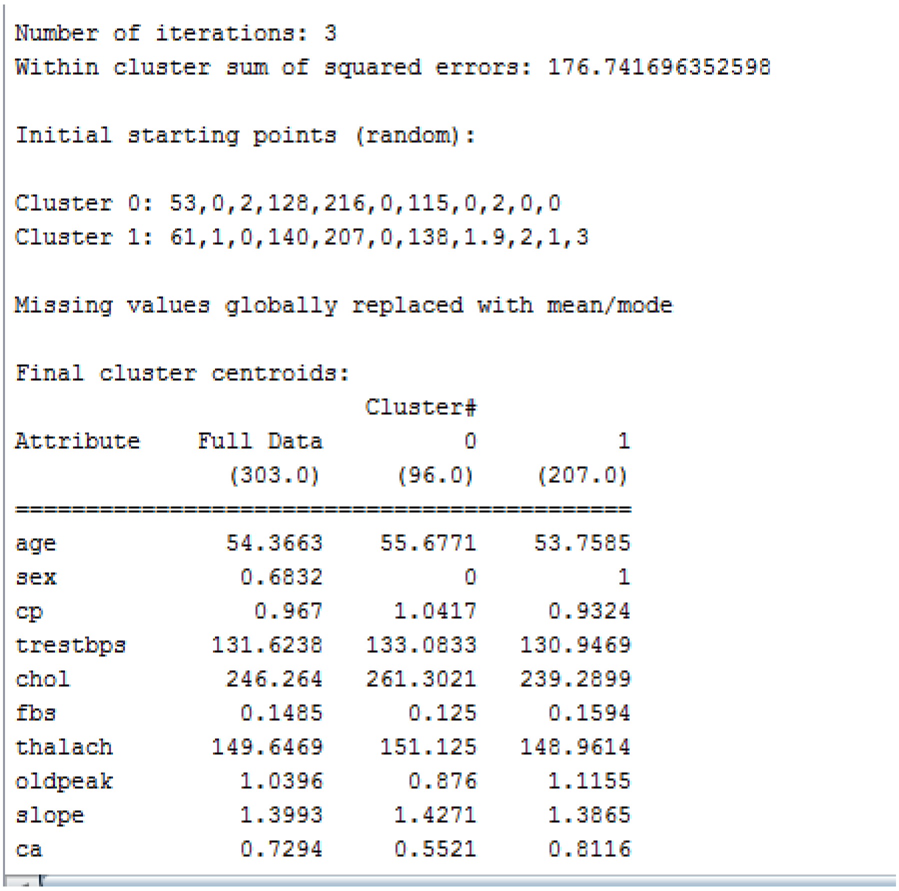
a. Implementing on wine dataset:



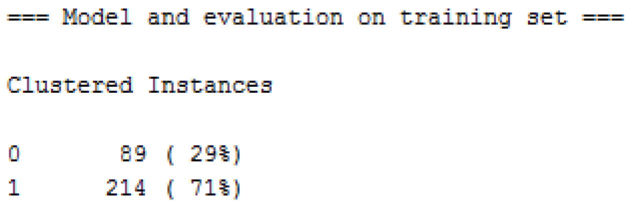
```
Fitted estimators (with ML estimates of variance):

Cluster: 0 Prior probability: 0.4255

Attribute: Closeness
Normal Distribution. Mean = 3.3166 StdDev = 0.3635

Cluster: 1 Prior probability: 0.5745

Attribute: Closeness
Normal Distribution. Mean = 1.2677 StdDev = 0.4785


Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      36 ( 39%)
1      56 ( 61%)
```



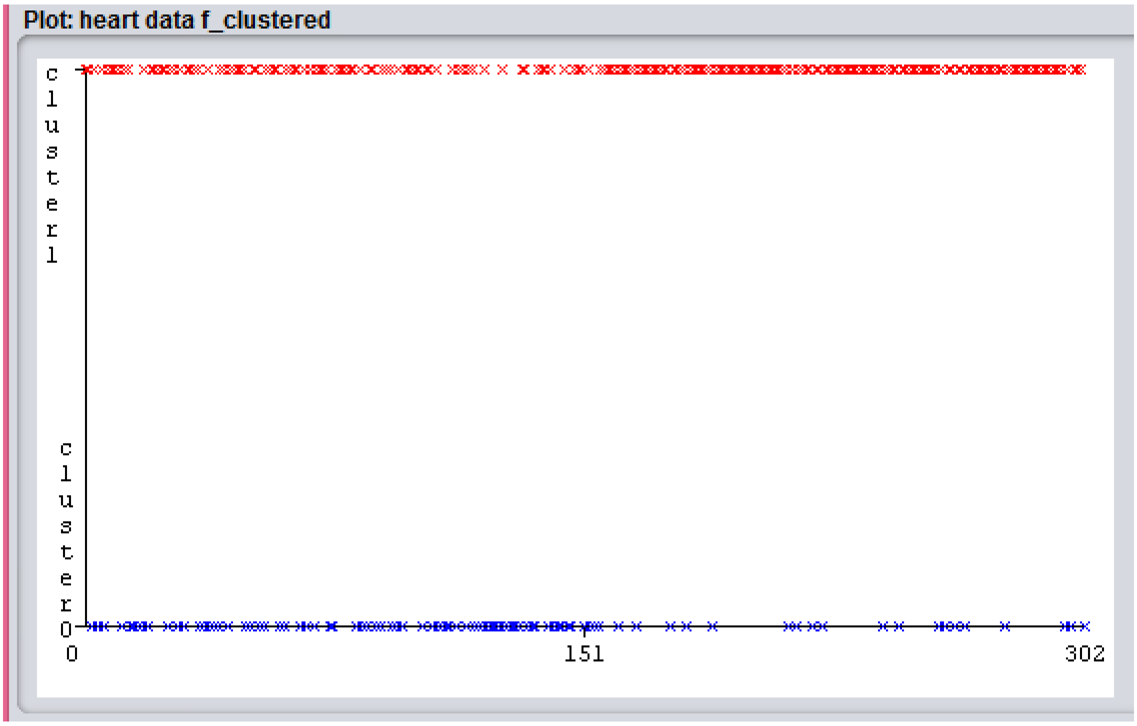
b. Implementing on heart dataset:



```
Number of iterations: 3
Within cluster sum of squared errors: 176.741696352598

Initial starting points (random):

Cluster 0: 53,0,2,128,216,0,115,0,2,0,0
Cluster 1: 61,1,0,140,207,0,138,1.9,2,1,3

Missing values globally replaced with mean/mode

Final cluster centroids:
                         Cluster#
Attribute    Full Data          0          1
               (303.0)     (96.0)    (207.0)
============================================
age            54.3663    55.6771    53.7585
sex             0.6832          0          1
cp               0.967     1.0417     0.9324
trestbps      131.6238   133.0833   130.9469
chol           246.264   261.3021   239.2899
fbs             0.1485      0.125     0.1594
thalach       149.6469    151.125   148.9614
oldpeak         1.0396      0.876     1.1155
slope           1.3993     1.4271     1.3865
ca              0.7294     0.5521     0.8116
```



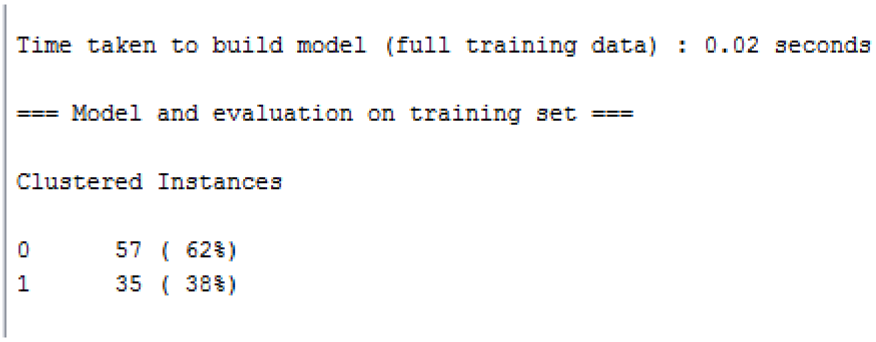
```
=== Model and evaluation on training set ===

Clustered Instances

0       89 ( 29%)
1      214 ( 71%)
```



2.1.1 Hierarchical Clustering:

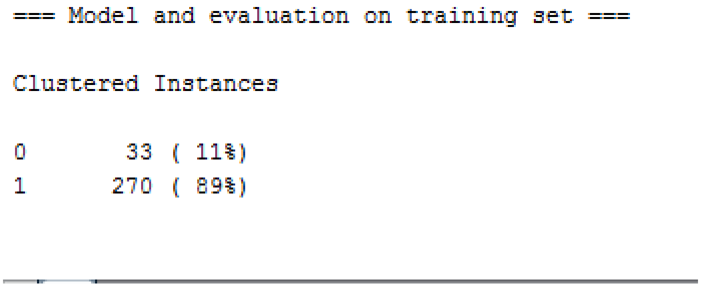
Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom.

a. **Implementing on wine dataset:**



```
Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       57 ( 62%)
1       35 ( 38%)
```

b. Implementing on heart Dataset:



```
=== Model and evaluation on training set ===

Clustered Instances

0        33 ( 11%)
1       270 ( 89%)
```

Conclusion

Thus we have studied and understood Weka ,an open source tool for data mining.