WEKA MANUAL

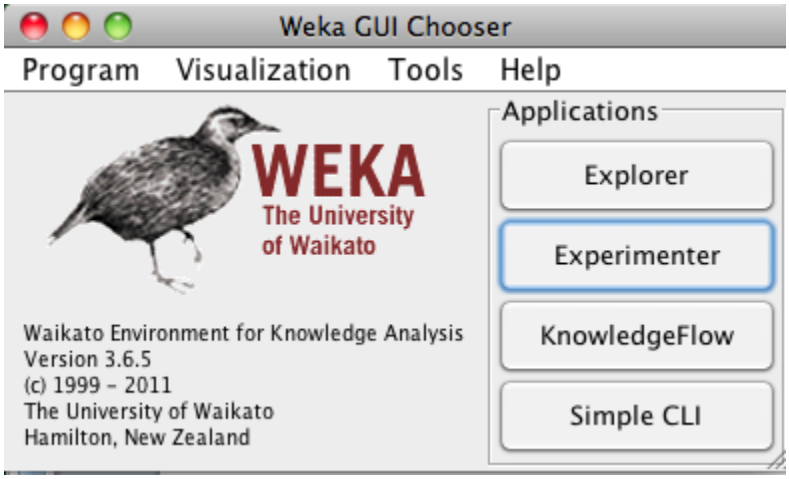
Introduction

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 re-processing, classification, clustering, regression and feature selection to name a few. The workflow of WEKA would be as follows:

Data → Pre-processing → Data Mining → Knowledge

Getting started with WEKA

Choose "WEKA 3.7.x" from Programs. The first interface that appears looks like the one given below.

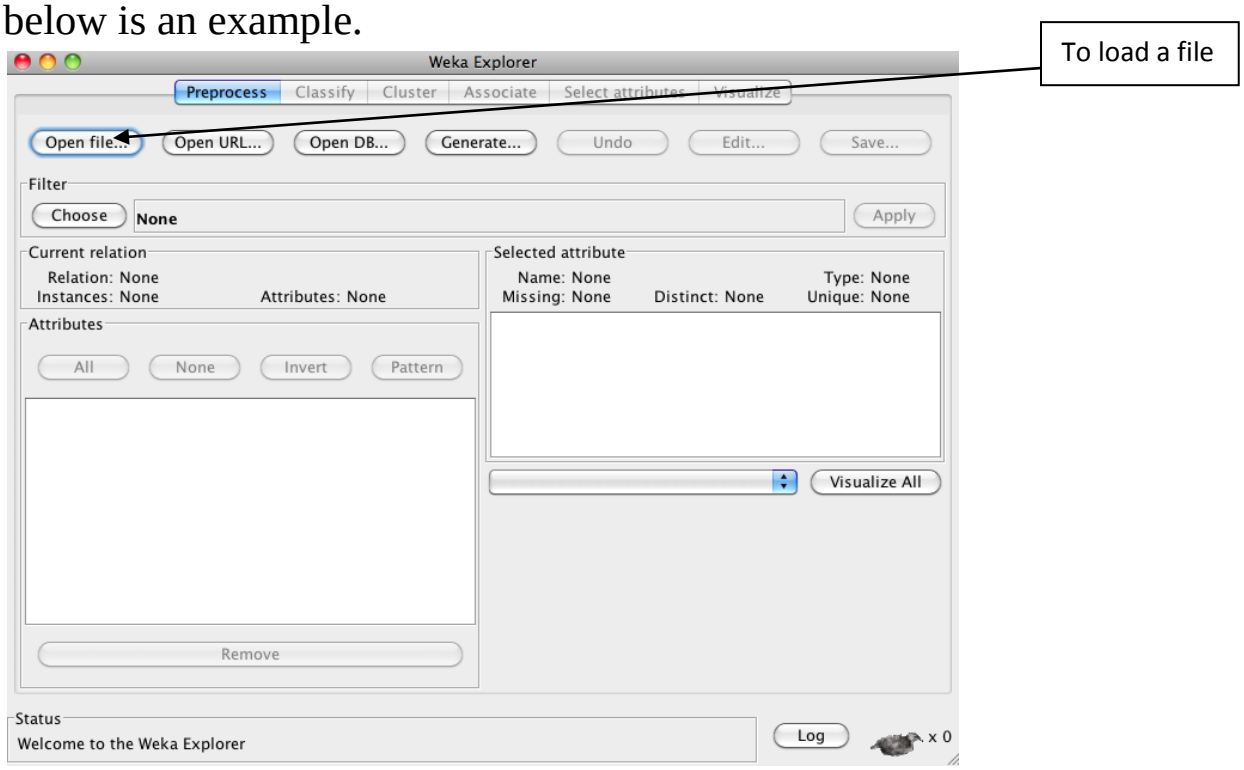


- **Explorer:** An environment for exploring data. It supports data preprocessing, attribute selection, learning and visualization
- **Experimenter:** An environment for performing experiments and conducting statistical tests between machine learning algorithms.

- **Knowledge Flow:** It is similar to Explorer but has a drag-and-drop interface. It gives a visual design of the KDD process.
- **Simple CLI:** Provides a simple command-line interface for executing WEKA commands.

WEKA Tools

- **Preprocessing Filters:** The data file needs to be loaded first. Given below is an example.



The supported data formats are **ARFF, CSV, C4.5 and binary**. Alternatively you could also import from URL or an SQL database. After loading the data, preprocessing filters could be used for **adding/removing attributes, discretization, Sampling, randomizing** etc.

- **Select attributes:** WEKA has a very flexible combination of search and evaluation methods for the dataset's attributes. Search methods

include **Best-first, Ranker, Genetic-search,** etc. Evaluation measures include **InformationGain, GainRatio, ReliefF**, etc.

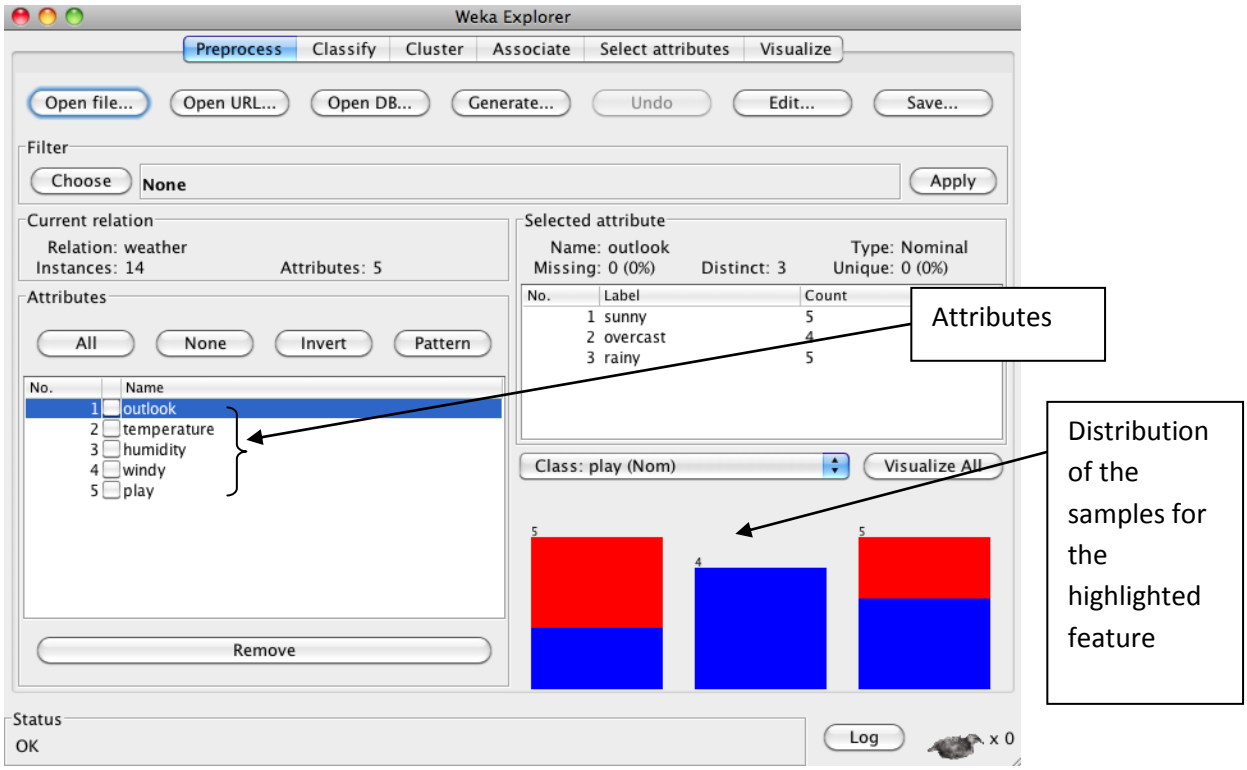
- **Classification:** The predicted target must be categorical. WEKA includes methods such as Decision Trees, Naïve Bayes and Neural Networks to name a few. Evaluation methods also include test data set and cross validation.

- **Clustering:** The learning process occurs from data clusters. Methods include k-means, Cobweb and FarthestFirst.

- **Regression:** The predicted target is continuous. Methods such as linear regression, Neural networks and Regression trees are included in the library.

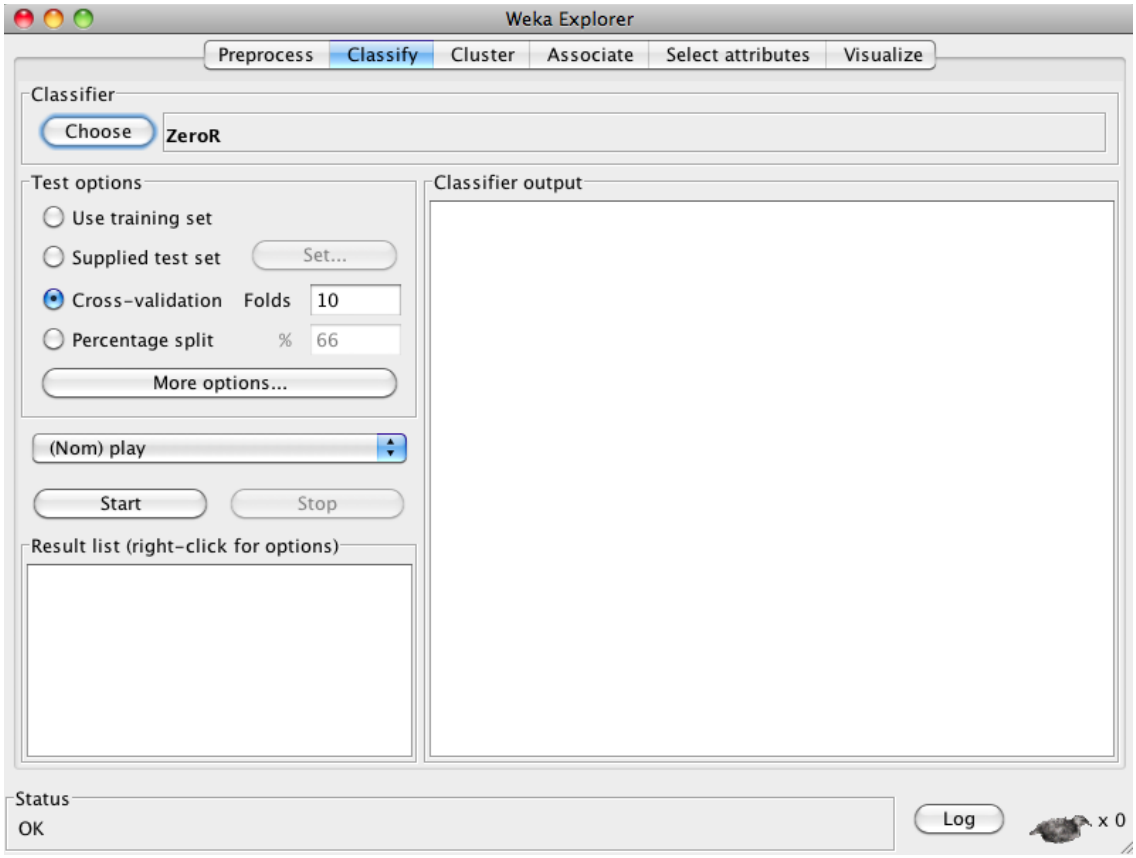
Exercise (Using built-in dataset)

1. Click Explorer on the first interface screen and load a dataset from the library. Given here is an illustration for the dataset 'weather.arff'.



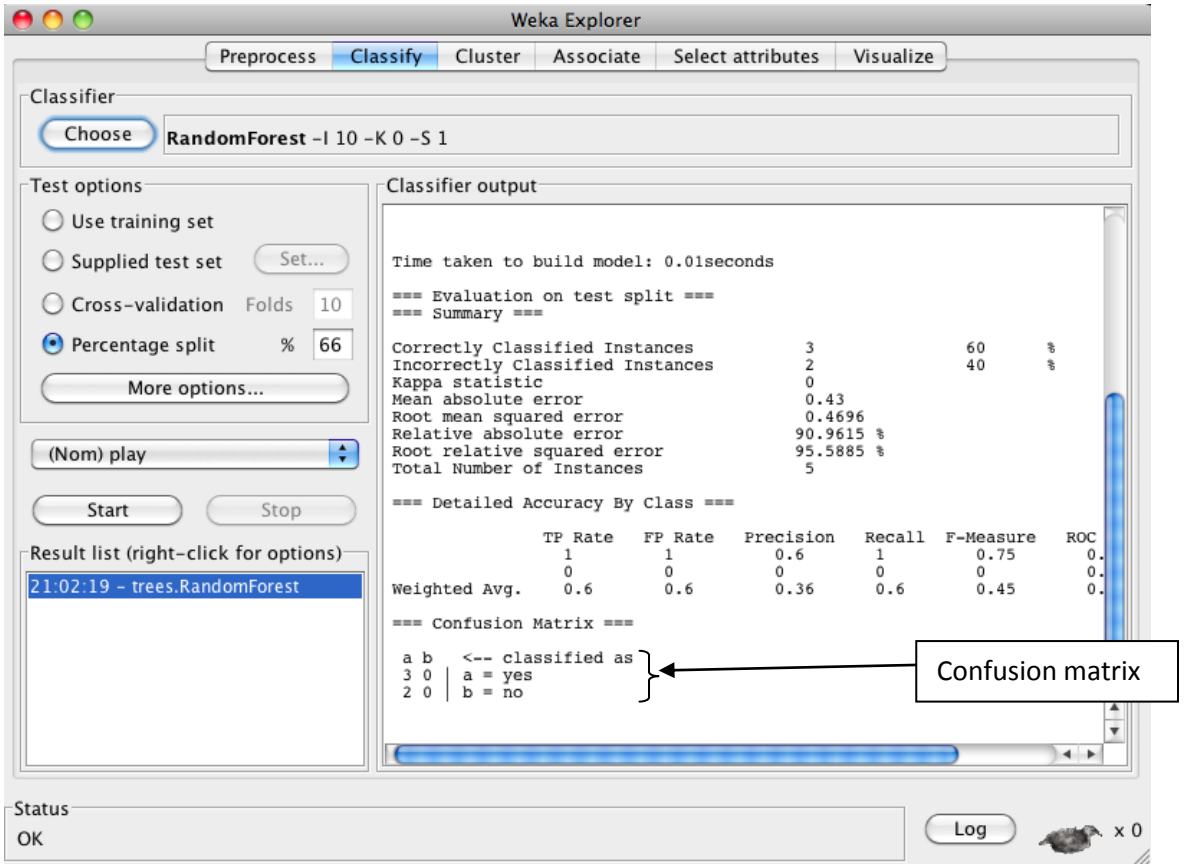
2. Click over each attribute to visualize the distribution of the samples for each of them. You can also visualize all of them at the same time by clicking the ‘Visualize all’ on the right pane.

3. Under the Classify tab, click ‘Choose’ and select a classifier from the drop-down menu. E.g.: ‘Decision Stump’



4. Once, a classifier is chosen, select percentage split and leave it with its default values. The default ratio is 66% for training and 34% for testing.

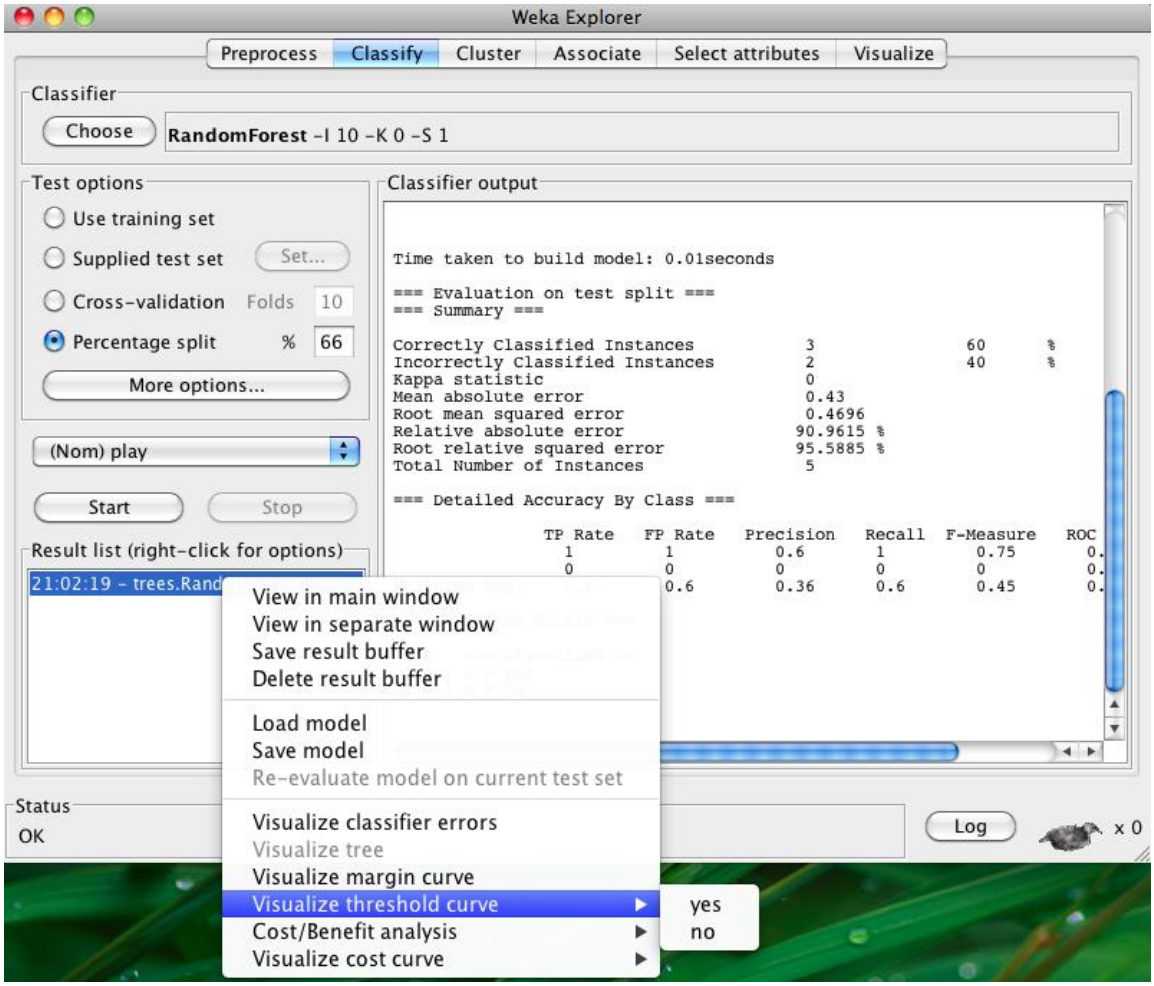
5. Click ‘Start’ to train and test the classifier. The interface will now like this:



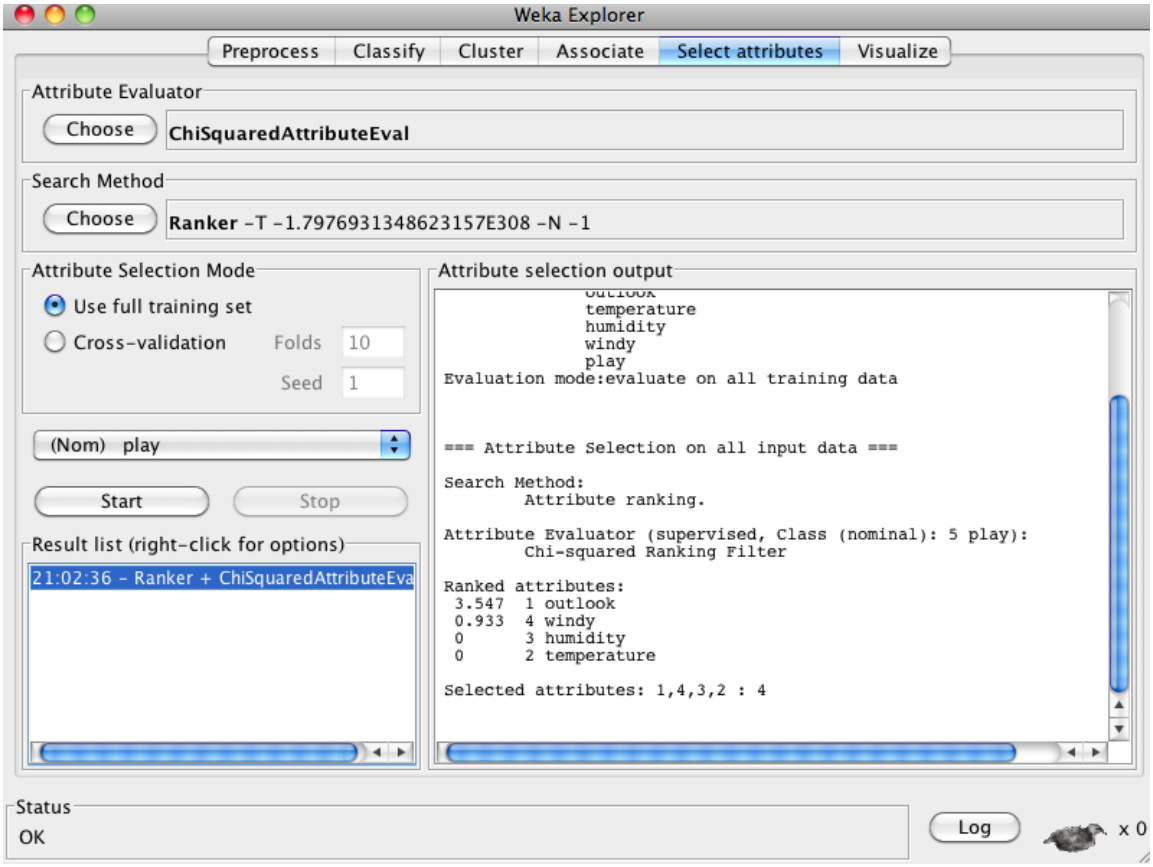
6. You could also try using ‘Crossvalidation’ method to train and test the data.

7. The right pane shows the results for training and testing. It also indicates the number of correctly classified and misclassified samples.

8. You could right click on the model generated and do various operations. You could also save the model if you wanted. Another performance measure is the ROC curve that can be viewed as shown in the next picture. Select ‘no’ in the option to view the curve.



9. Click on the ‘Select Attributes’ tab and to analyze the attributes. A number of ‘Attribute Evaluator’ and ‘Search methods’ can be combined to gain insight about the attributes. Given below is an example.



10. Click on the Visualize tab to see the pair wise relationship of the attributes.

Performance Analysis

Once the model has been trained and tested, we need to measure the performance of the model. For this purpose we used three measures namely: precision, recall and accuracy.

Precision (P) = tp/(tp+fp)

Recall (R) = tp/(tp+fn)

Accuracy (A) = (tp+tn)/Total # samples

Where tp, fp, tn and fn are true positive, false positive, true negative and false negative respectively.

Deliverables:

Use any preexisting dataset from the WEKA library. Choose any classifier and perform all the steps mentioned above. In your HW, please summarize the following:

1. Classifier used
2. Test Options used
3. Confusion Matrix in the form of a table
4. Figure of the ROC curve
5. The top 2 ranked attributes when you choose the Attribute Evaluator as ‘ChiSquaredAttributeEval’ and Search Method as ‘Ranker’.
6. P, R and A measures

Acknowledgement: This handout is a guide for WEKA for EECS 730 students at KU only. Some information above is taken from a few related internet sources.