INFSCI 2725: Data Analytics

Assignment 7

Jinrong Liu, Fengxi Liu

Objects:

* Data that should be analyzed: Iris data ( <http://archive.ics.uci.edu/ml/datasets/Iris> ) and Congressional Voting Records ( <http://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records> )
* Try different learning algorithms and their parameters then check and report their classification accuracy for each data set.
* Try to improve the accuracy by using different learning methods for each data set.
* Report the best classification accuracy that you have been able to achieve for each of the two data sets along with the methods that have the best accuracy.

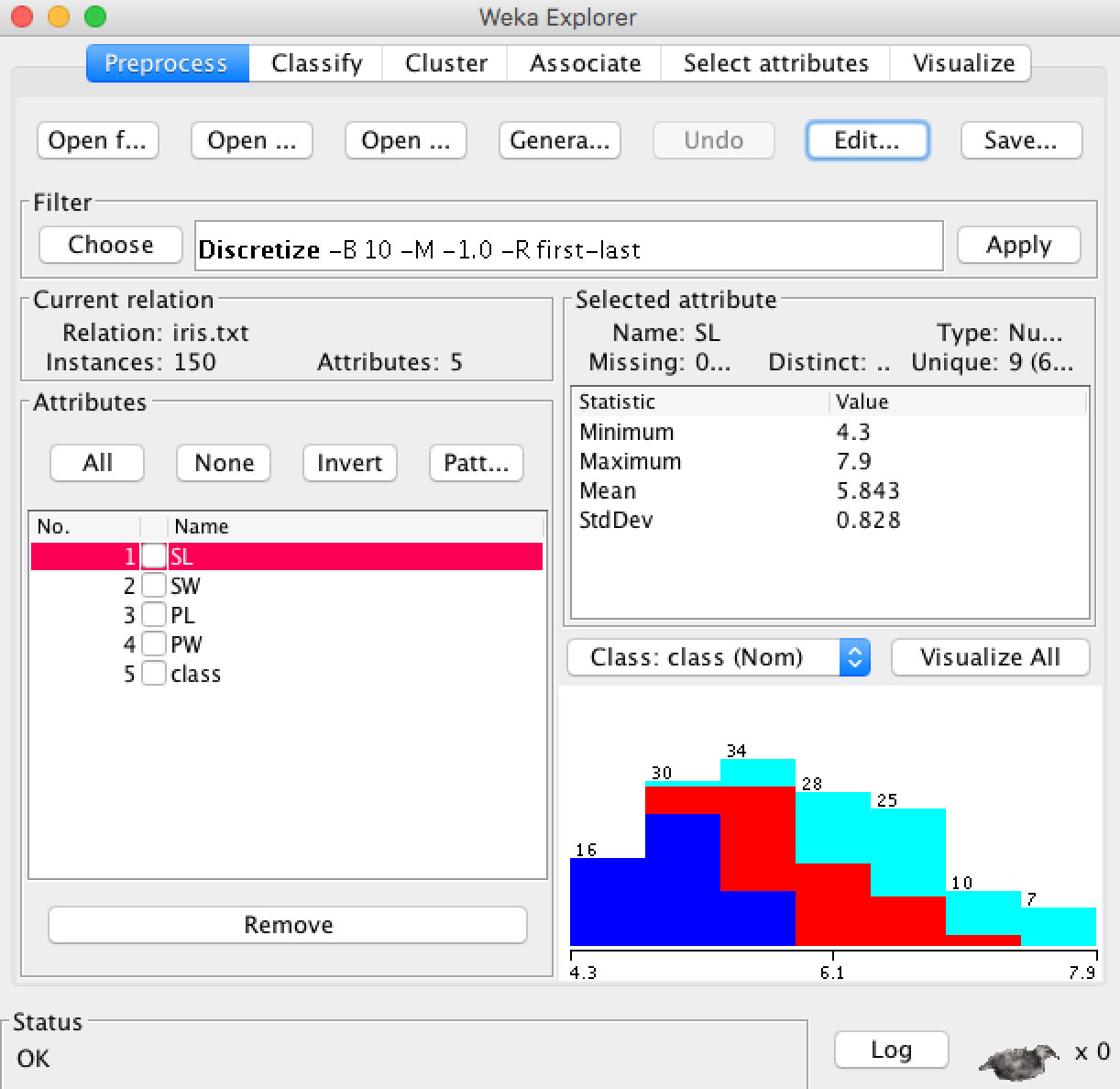
Analytical tool: Weka

1. Iris data set

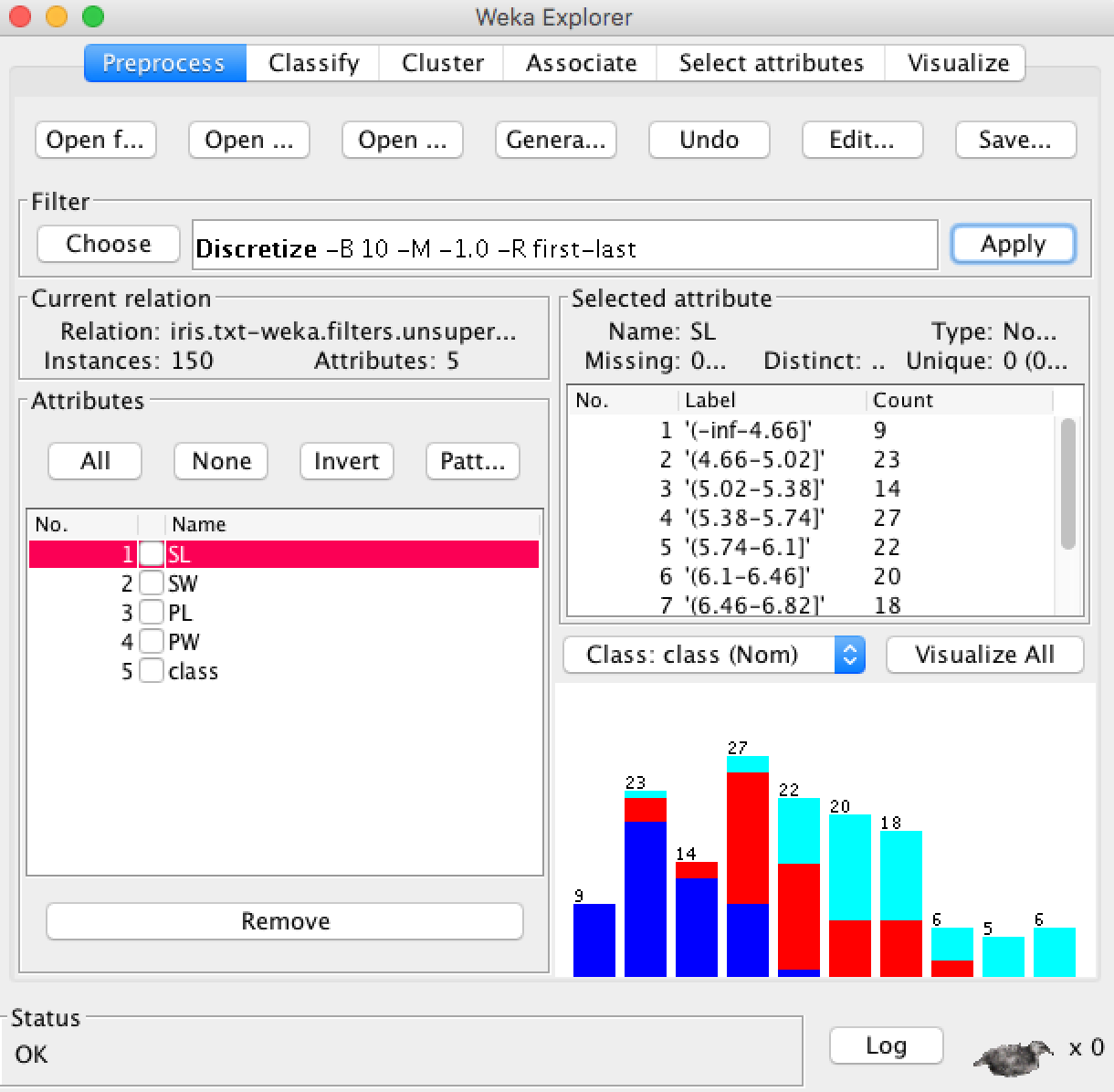
* Step 1: Discretize the data

We import the Iris data into Weka and discretize the data firstly.

Before we discretized the data:

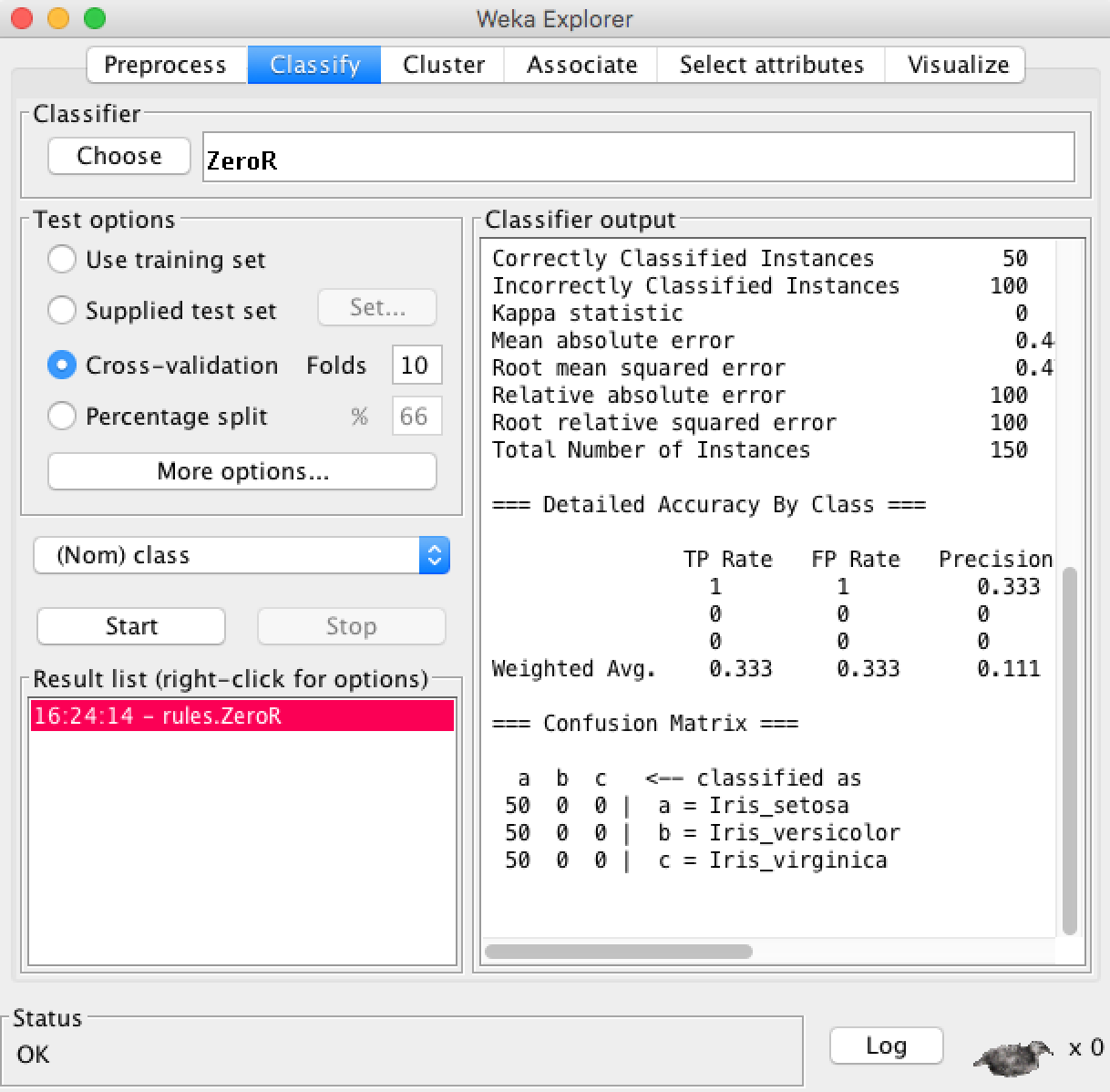


After we discretized the data:



* Step 2: Select and run and algorithm

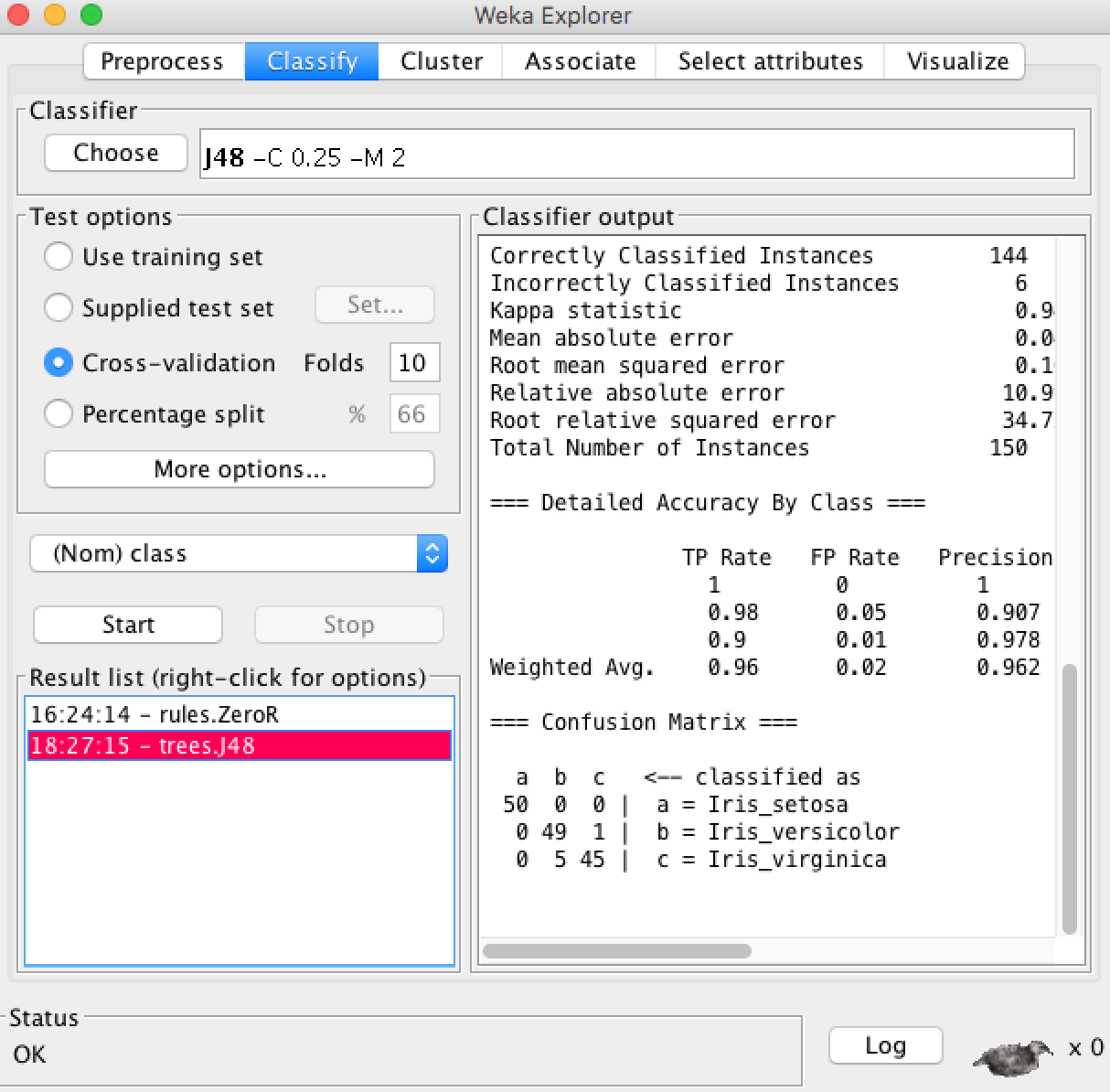
By default, the algorithm is set as “ZeroR”, the result is as follows:



The **ZeroR** algorithm selects the majority class in the dataset, in this case, all three species of iris are equally present in the data, so it picks the first one: iris setosa. It uses this to make all predictions. This is the baseline for the dataset and the measure by which all algorithms can be compared. The result is 33%, which means that each of the three classes are equally represented, assigning one of the three to each prediction results in 33% classification accuracy.

We note that the test options use Cross Validation by default with 10 folds. This means that the dataset is split into 10 parts, the first 9 are used to train the algorithm, and the 10th is used to access the algorithm. This process is repeated allowing each of the 10 parts of the split dataset a chance to be held out test set.

Then we tried another algorithm which is “**J48**” (C4.8 algorithm in Java), the result is as follows:



This algorithm was also run with 10 folds cross validation, which means that it was given an opportunity to make a prediction for each instance of the dataset.

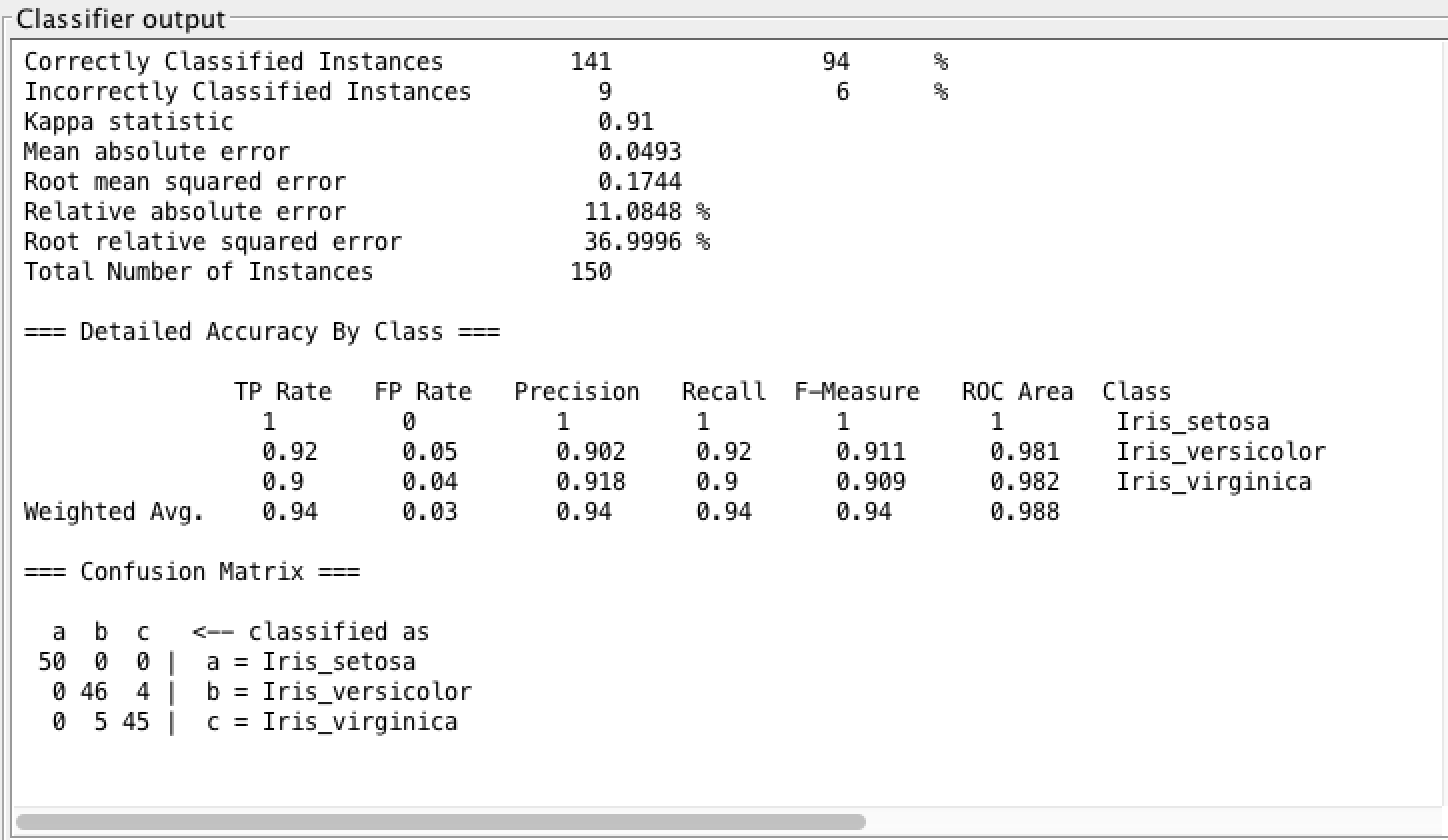
Using J48 algorithm, the classification accuracy should be 144/150, which is 96%, it is a lot better than the baseline of 33%.

From Confusion Matrix, we can see a table of actual classes compared predicted classes. For iris\_setosa, the classification worked well; for iris\_versicolor, there was 1 error where iris\_versicolor was classified as iris\_virginica; for iris\_virginica, there were 5 cases that the iris\_virginica were classified as iris\_versicolor.

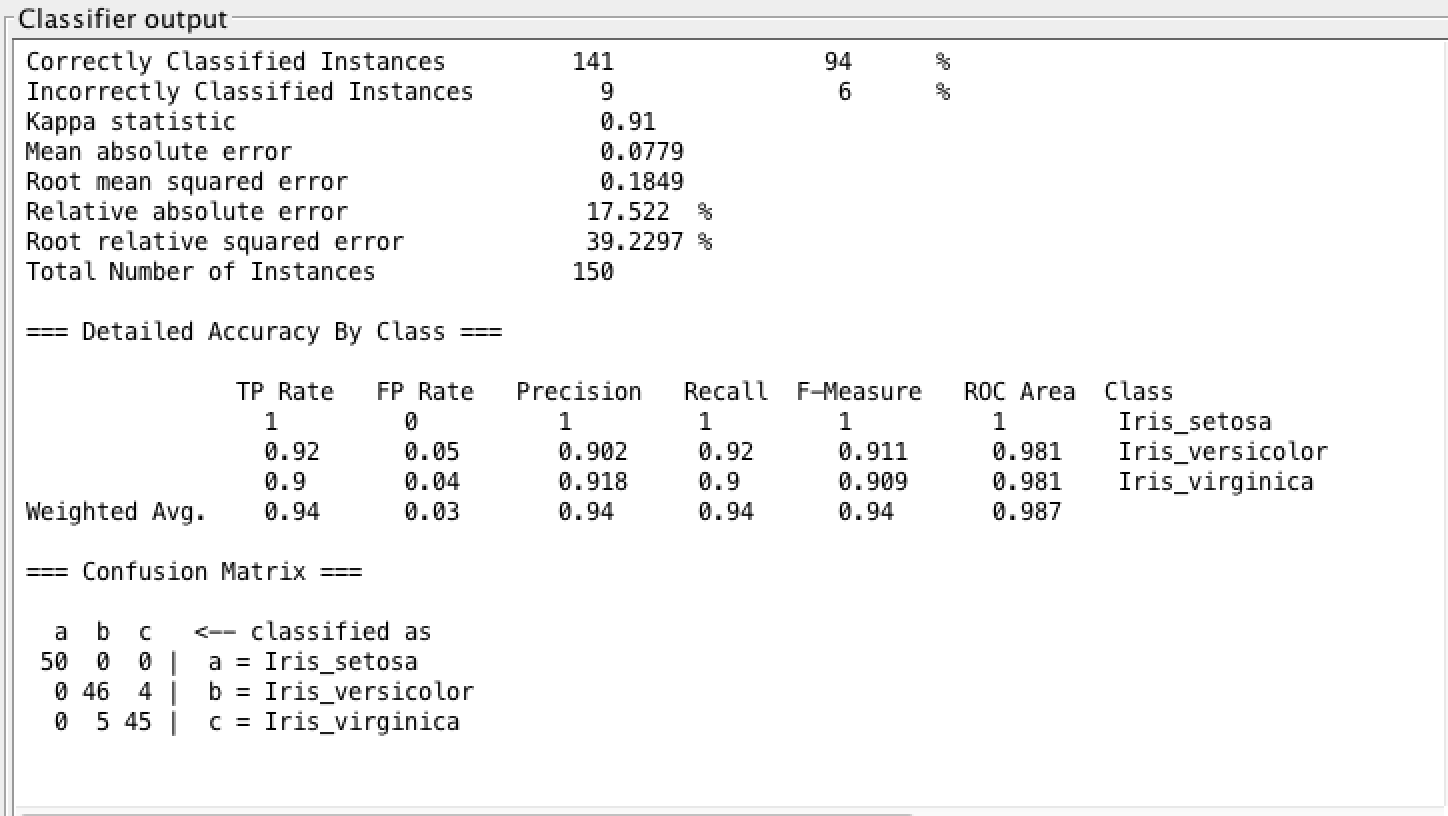
We briefly tried other algorithms like **BayesNet** and **KStar**.

Their Classifier outputs are as follows:

BayesNet:



KStar:



It is interesting that these two algorithms showed almost the same result according to the classification accuracy and confusion matrix.

Comparing all the algorithms that we have tried, the J48 algorithm showed the best classification accuracy which is 96%.

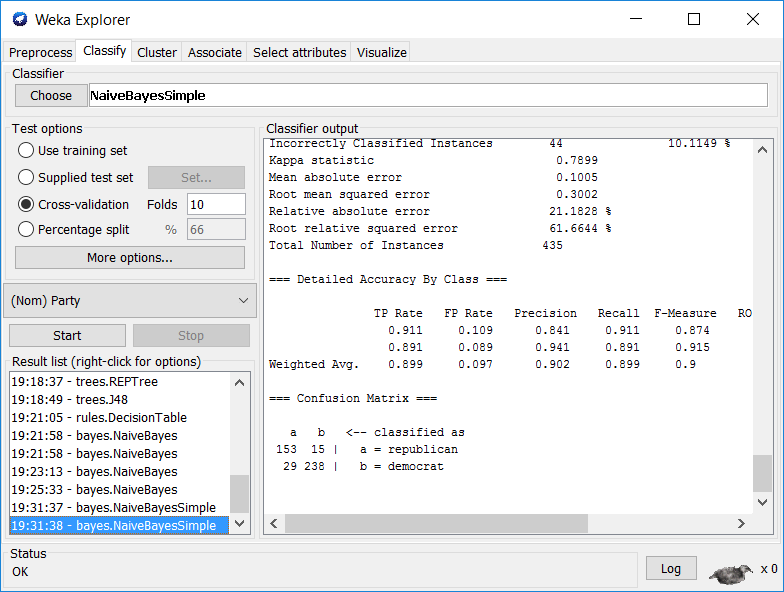
1. Congressional Voting Records data set.

Step 1: Data clean up and discretize

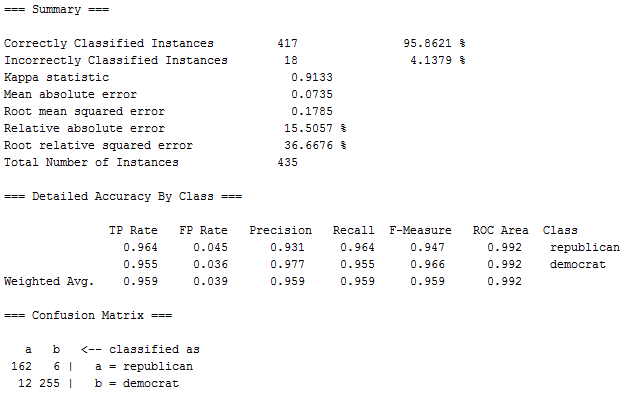
We used filter/unsupervised/discretize in WEKA GUI to do the data discretize.

Step 2: Clssification

We use “10 –folds “cross validation to check the classification accuracy of different algorithms. We use the rate of correctly classified instances as a measurable indicator of classification accuracy.



After trying different algorithms, we conclude that except several algorithms (such as ZeroR 61%), the correctly classified instances rate is mainly ranged between 88% -95%. In all the algorithms we have tested, RandomForest(trees) got the highest correctly classified rate.(95.8621%) Here is the result of Random Forrest classification.



Instances that are correctly classified are 417 out of 500, which means the correctly classified rate is 95.8621%.

For details, we can know from the confusion matrix. A represents the republican while b represents democrat. There are 174 republicans while 162 of them are correctly classified. 12 of the republican are wrongly classified as democrat. And there are 261 democrats while 255 of them are correctly classified and 6 of them are wrongly classified as republican. Confusion matrix gives us more details about the classification. And we can conclude that Random Forrest is a successful classification model for the data set of congressmen.