

INFSCI 2725: Data Analytics

Assignment 7

Tong Wei (TOW6) JinRong Liu (JIL181) FengXi Liu (FEL34) Zhenyu Peng(ZHP6)

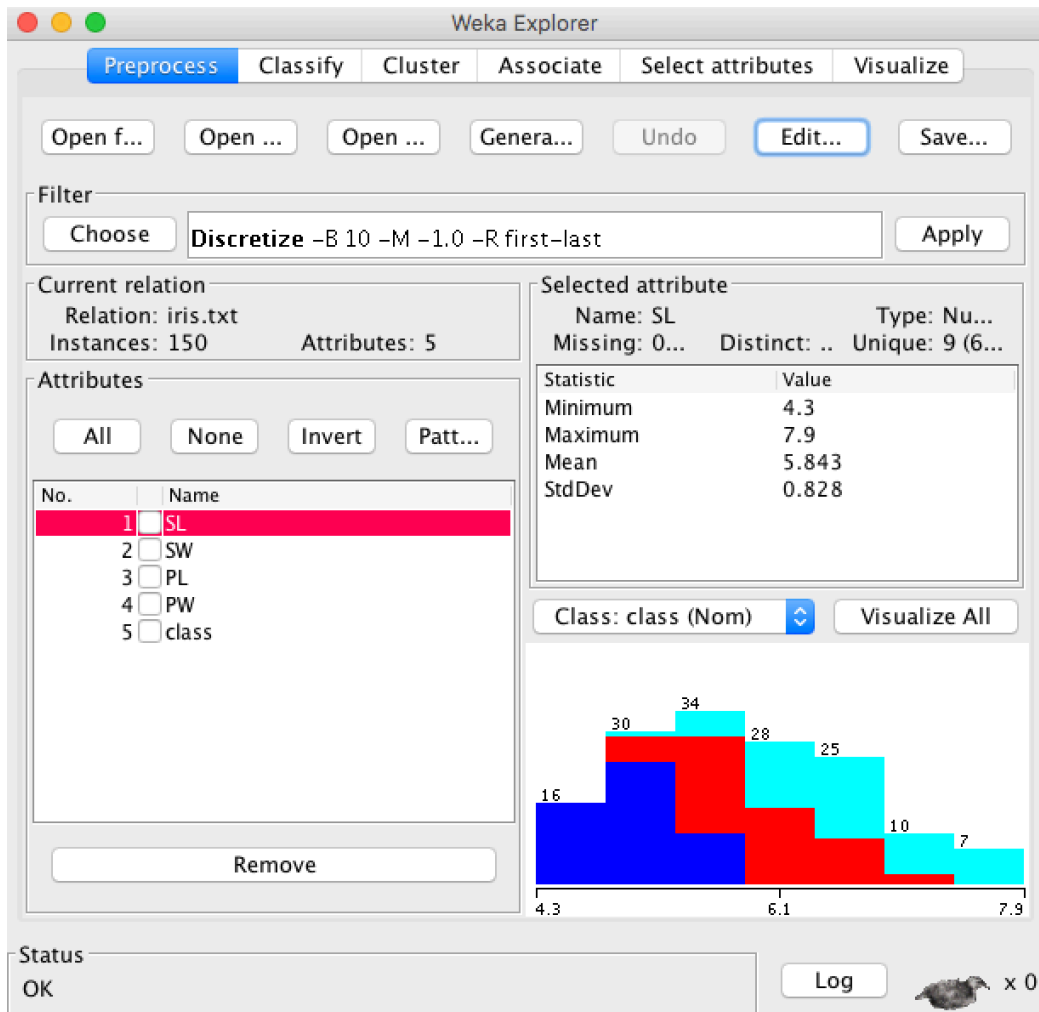
Objects:

- 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.

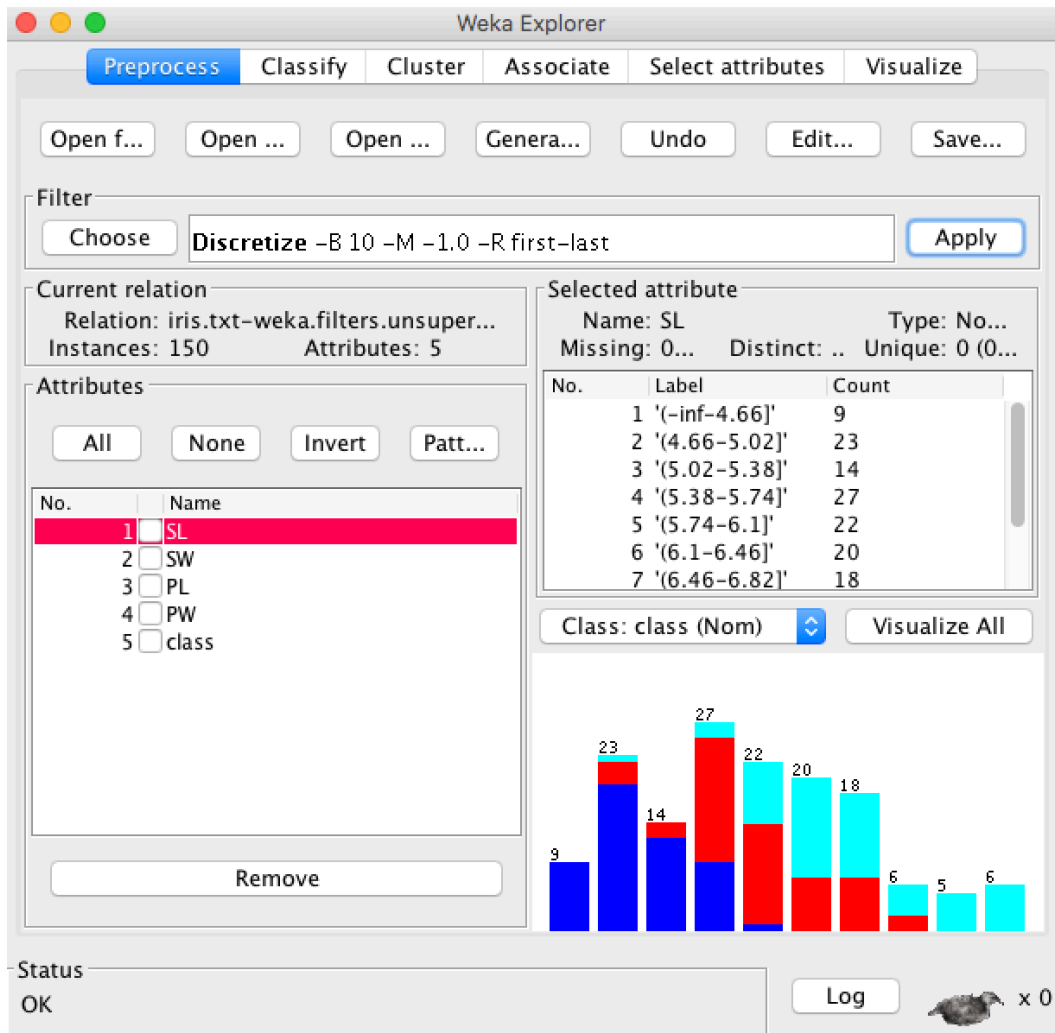
Analytical tool: Weka

1. Iris data set

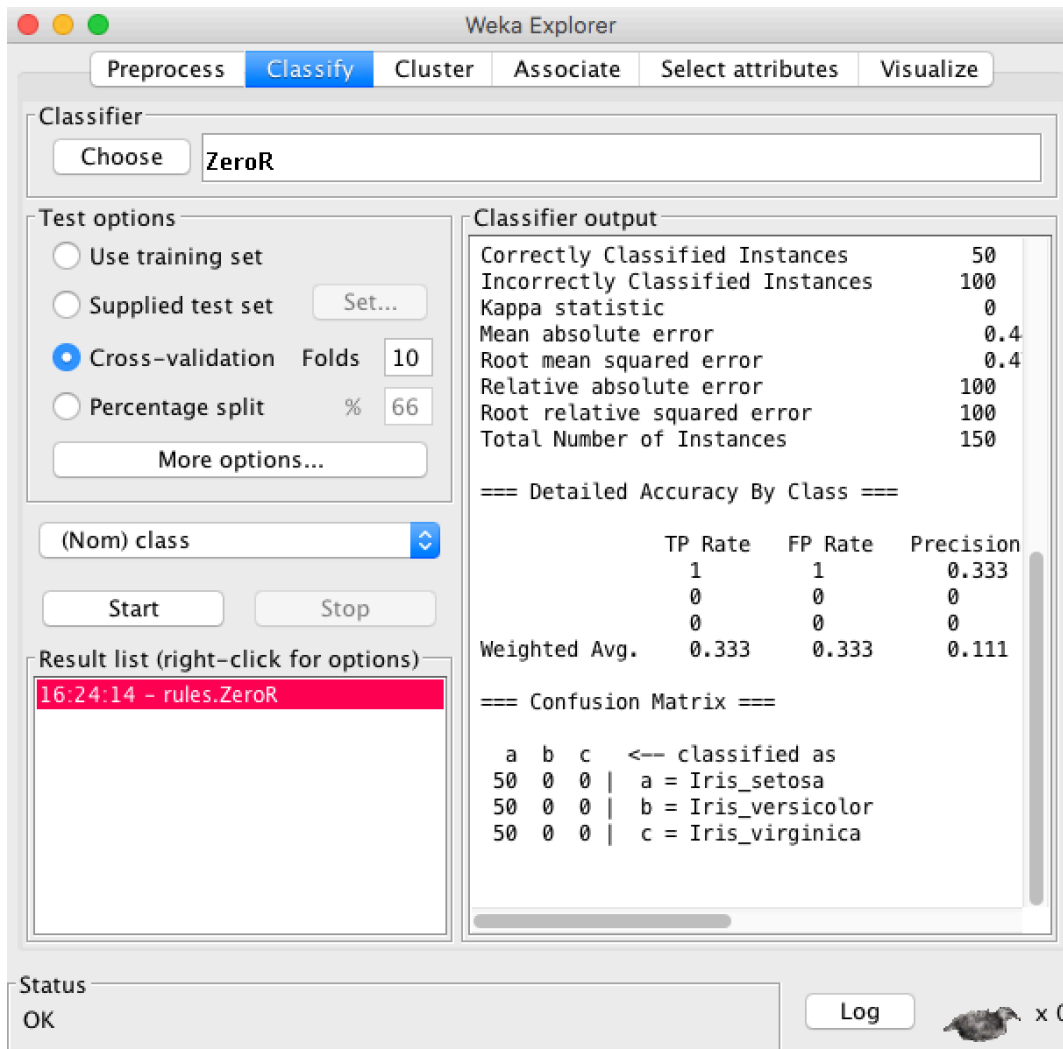
- Step 1: Discretize the data
We import the Iris data into Weka and discretize the data firstly.
Before we discretized the data:



After we discretized the data:



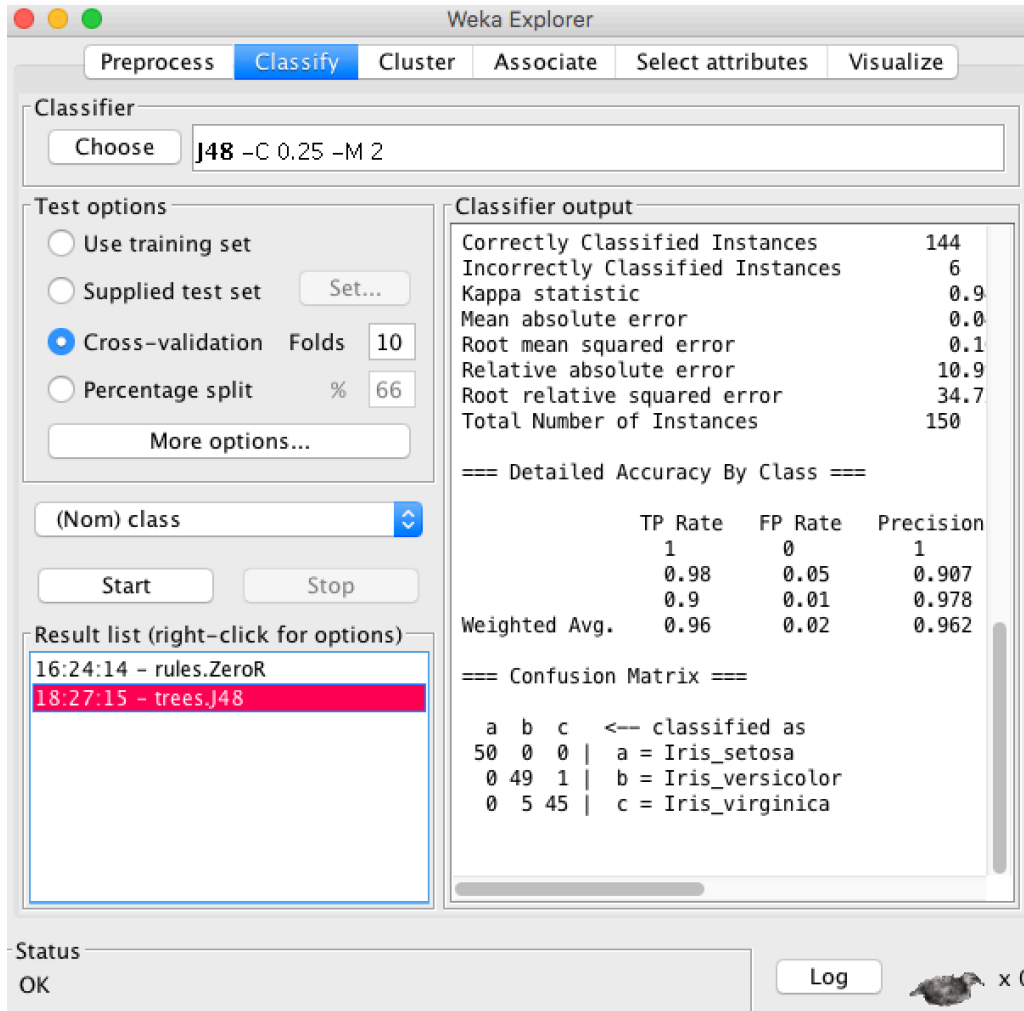
- Step 2: Select and run an algorithm
By default, the algorithm is set as “ZeroR”, the result is as follows:



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:



The screenshot shows the Weka Explorer application window. The 'Classify' tab is selected. The classifier chosen is 'J48 -C 0.25 -M 2'. The test options are set to 'Cross-validation' with 'Folds' set to 10. The classifier output is displayed on the right, showing various performance metrics and a confusion matrix.

Classifier
Choose J48 -C 0.25 -M 2

Test options
☐ Use training set
☐ Supplied test set Set...
☒ Cross-validation Folds 10
☐ Percentage split % 66
More options...

(Nom) class

Start Stop

Result list (right-click for options)
16:24:14 - rules.ZeroR
18:27:15 - trees.J48

Classifier output

Correctly Classified Instances 144
Incorrectly Classified Instances 6
Kappa statistic 0.9
Mean absolute error 0.0
Root mean squared error 0.1
Relative absolute error 10.9
Root relative squared error 34.7
Total Number of Instances 150

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision
1	1	0	1
0.98	0.05	0.907	
0.9	0.01	0.978	
Weighted Avg.	0.96	0.02	0.962

=== Confusion Matrix ===

a	b	c	<-- classified as
50	0	0	a = Iris_setosa
0	49	1	b = Iris_versicolor
0	5	45	c = Iris_virginica

Status
OK

Log x 0

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:

Classifier output							
Correctly Classified Instances	141	94	%				
Incorrectly Classified Instances	9	6	%				
Kappa statistic	0.91						
Mean absolute error	0.0493						
Root mean squared error	0.1744						
Relative absolute error	11.0848 %						
Root relative squared error	36.9996 %						
Total Number of Instances	150						
=== Detailed Accuracy By Class ===							
	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	0	1	1	1	1	Iris_setosa
	0.92	0.05	0.902	0.92	0.911	0.981	Iris_versicolor
	0.9	0.04	0.918	0.9	0.909	0.982	Iris_virginica
Weighted Avg.	0.94	0.03	0.94	0.94	0.94	0.988	
=== Confusion Matrix ===							
a	b	c	<-- classified as				
50	0	0	a = Iris_setosa				
0	46	4	b = Iris_versicolor				
0	5	45	c = Iris_virginica				

KStar:

Classifier output							
Correctly Classified Instances	141	94	%				
Incorrectly Classified Instances	9	6	%				
Kappa statistic	0.91						
Mean absolute error	0.0779						
Root mean squared error	0.1849						
Relative absolute error	17.522 %						
Root relative squared error	39.2297 %						
Total Number of Instances	150						
=== Detailed Accuracy By Class ===							
	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	0	1	1	1	1	Iris_setosa
	0.92	0.05	0.902	0.92	0.911	0.981	Iris_versicolor
	0.9	0.04	0.918	0.9	0.909	0.981	Iris_virginica
Weighted Avg.	0.94	0.03	0.94	0.94	0.94	0.987	
=== Confusion Matrix ===							
a	b	c	<-- classified as				
50	0	0	a = Iris_setosa				
0	46	4	b = Iris_versicolor				
0	5	45	c = Iris_virginica				

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%.

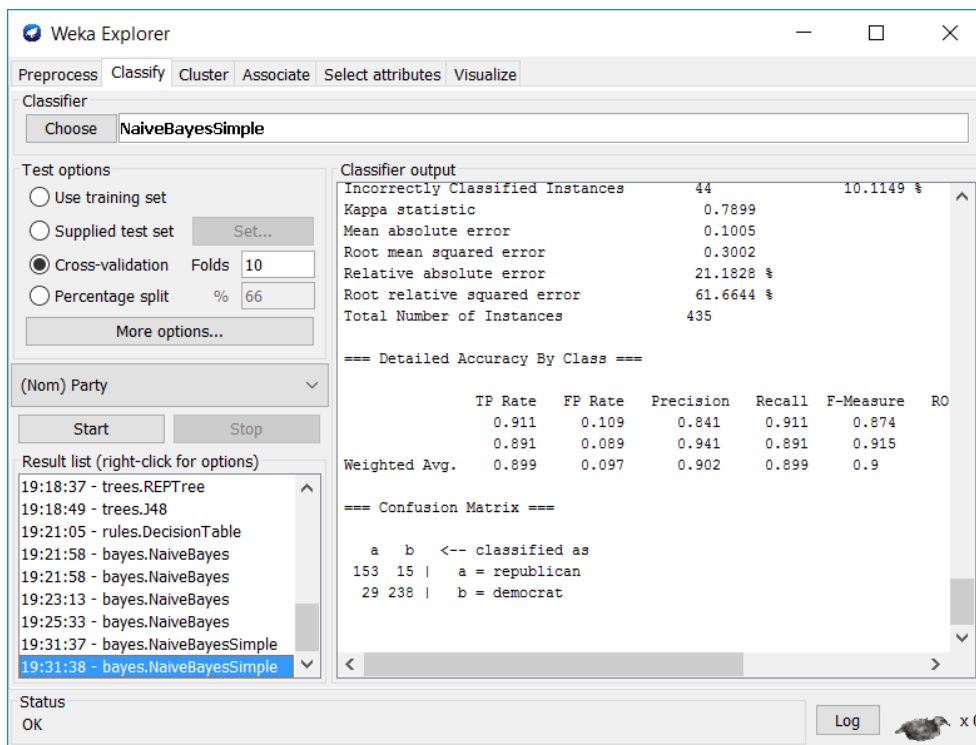
2. Congressional Voting Records data set.

Step 1: Data clean up and discretize

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

Step 2: Classification

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.



The screenshot shows the Weka Explorer window with the 'Classify' tab selected. The classifier chosen is 'NaiveBayesSimple'. The test options are set to 'Cross-validation' with 'Folds' set to '10'. The classifier output is displayed, showing various performance metrics and a confusion matrix.

Classifier output

Metric	Value	Percentage
Incorrectly Classified Instances	44	10.1149 %
Kappa statistic	0.7899	
Mean absolute error	0.1005	
Root mean squared error	0.3002	
Relative absolute error	21.1828	%
Root relative squared error	61.6644	%
Total Number of Instances	435	

Detailed Accuracy By Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	RO
0.911	0.109	0.841	0.911	0.874		
0.891	0.089	0.941	0.891	0.915		
Weighted Avg.	0.899	0.097	0.902	0.899	0.9	

Confusion Matrix

```
a  b  <-- classified as
153 15 | a = republican
29 238 | b = democrat
```

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.

```

=== Summary ===

Correctly Classified Instances      417          95.8621 %
Incorrectly Classified Instances    18           4.1379 %
Kappa statistic                     0.9133
Mean absolute error                 0.0735
Root mean squared error            0.1785
Relative absolute error             15.5057 %
Root relative squared error        36.6676 %
Total Number of Instances          435

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
      0.964    0.045    0.931    0.964    0.947    0.992    republican
      0.955    0.036    0.977    0.955    0.966    0.992    democrat
Weighted Avg.   0.959    0.039    0.959    0.959    0.959    0.992

=== Confusion Matrix ===

  a    b  <-- classified as
162    6 |  a = republican
 12 255 |  b = democrat

```

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.