Practical Data Mining

COMP-321B

Tutorial 1: Introduction to the WEKA Explorer

Gabi Schmidberger
Mark Hall
Richard Kirkby

July 12, 2006

1 Setting up your Environment

Before you can start WEKA the PATH must contain a reference to a JAVA environment and the CLASSPATH contain the 'weka.jar' file.

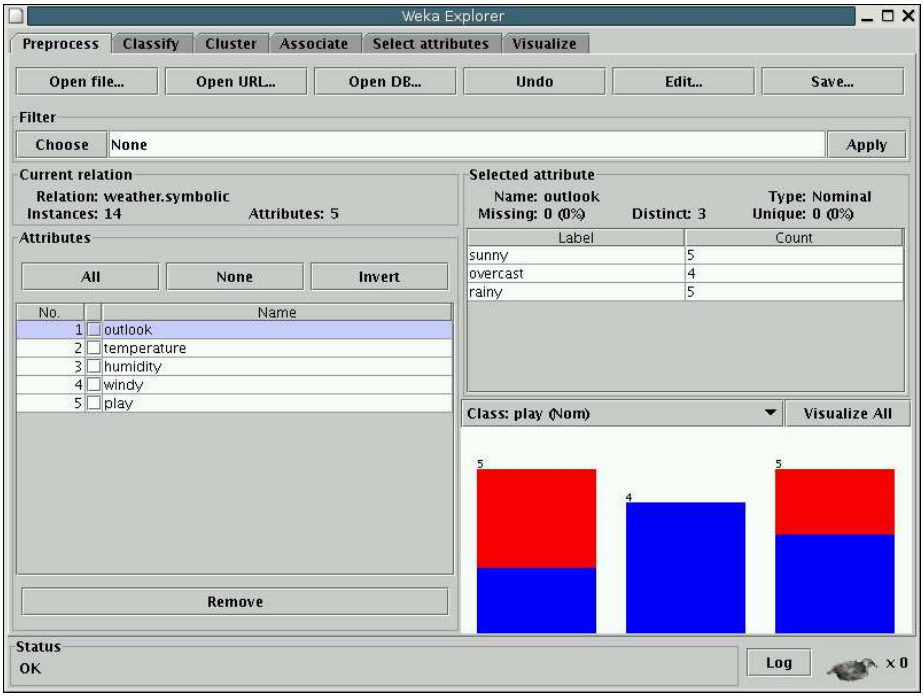
```
export PATH=/home/ml/jdk/jdk1.5.0_07/bin:$PATH
export CLASSPATH=/home/ml/321/weka.jar:$CLASSPATH
```

2 Start up the WEKA-Explorer

Open a command window and enter

```
java -Xmx500M weka.gui.explorer.Explorer
```

You will have to wait a few seconds for the **WEKA Explorer** to open up.



3 Data Set Files

All Files used in this tutorial are in the folder:

```
/home/ml/321/Tutorial1
```

4 The Different Panels in WEKA

After you have started up the WEKA Explorer you can see that the user interface is split into six panels. On top of the window are the tabs of these panels and the **Preprocess** panel is the current one.

This tutorial will explain only three of these six panels, the **Preprocess** panel, the **Classify** panel and the **Visualize** panel. The last three panels are similar to the **Classify** panel and any special functions of these panels will be explained in the introductions of the tutorials that use them.

This is an overview of the functions that these three panels perform.

Preprocess Here a data set can be loaded. After the data set is loaded the panel displays certain information about this data set. The data can also be changed either by editing it in the data set editor or by applying a filter. The changed data set can be saved. As an alternative to loading a pre-existing data set, an artificial one can be created by using a generator.

Classify From this panel the classification methods can be started. Several options for the classification process can be set and the results of the classification can be viewed. The training data set used for classification is the one loaded (or generated) in the 'Preprocess' panel.

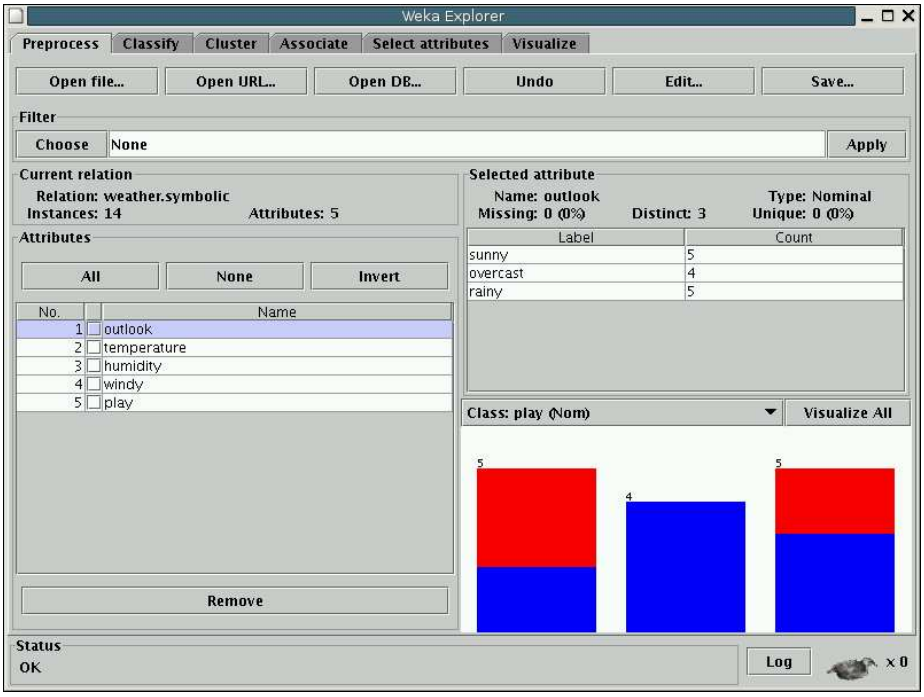
Visualize The data set loaded in the 'Preprocess' panel can be visualized in two dimensional pixel plots. The user selects the attributes for the x-axis and the y-axis of the plot.

5 The Preprocess Panel

5.1 Load a Data Set

The 'Preprocess' panel is the panel opened after starting the WEKA Explorer. Before changing to any of the other panels the Explorer must have a data set to work with. To load up a data set, click on the **Open file...** button in the top left corner of the window. Inside the data folder you will find the file named 'weather.nominal.arff'. Open this file.

After you have done this you will see:



The weather data is a small data set with only 14 examples. Instead of examples we will from now on use the term instances. The instances of the weather data set have 5 attributes, which have the names 'outlook', 'temperature', 'humidity', 'windy' and 'play'. If you click on the name of an attribute in the left panel the right panels will show more information about this attribute. You see in the right panel which values the attribute can have and how many times an instance in this data set has this value. Below this information is a small histogram drawn.

All attributes of this data set are nominal, which means the values can be one of a defined set of values. The data in one instance describes a weather forecast for a particular day and the decision to play golf or not on that day. The last attribute 'play' is the class attribute, it classifies the instance. The values can be 'yes' or 'no'. 'Yes' means the weather conditions are OK to play golf and 'no' means they are not OK.

5.2 Exercises

To get more familiar with the functions discussed so far, please do the following two exercises. The solutions to these and any following exercises in this tutorial are on the last page of this text.

Ex. 1: What are the values that the attribute 'temperature' can have?

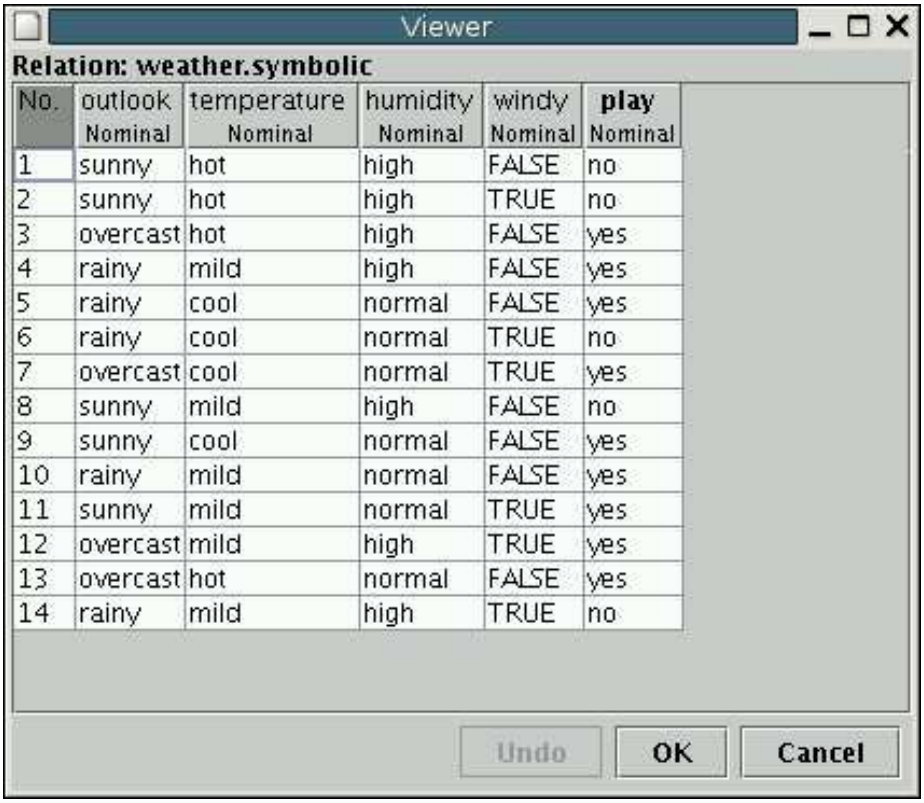
Ex. 2: Load a new data set: Press the 'Open file' button and select the file 'iris.arff'. How many instances does this data set have? How many at-

tributes does this data set have? What are the possible values of the attribute petallength?

5.3 The Data Set Editor

Please load the file 'weather.nominal.arff' again.

Click on the **Edit...** button from the row of buttons at the top of the window and a new window opens. The window is called **Viewer**, and lists all instances of the weather data represented in a table (see below).



5.3.1 Exercises

Ex. 3: View the weather data with the data set editor. What is the class value of the instance number 8.

Ex. 4: Is 'No.' (column one in 'Viewer') an attribute of the weather data set?

Ex. 5: Load the 'iris' data again. How many numeric and how many nominal attributes does this data set have?

5.4 Apply a Filter

In WEKA filters are methods that change data sets. WEKA has several filters for different tasks implemented. We want to use the filter that removes attibutes. This filter is called **Remove**. Its full name is:

weka.filters.unsupervised.attribute.Remove

Filters have names that are organized in a hierachical structure and 'weka' is the root of the structure. If you click the **Choose** button a hierachical editor opens and you can select the filter following the path of its full name. After you have selected the 'Remove' filter the text 'Remove' appears in the field right beside the **Choose** button.

Click on the field right beside the **Choose** button and a window opens. The window contains a short explanation of the filter and two fields in which the options of the filter can be set. The first option is a list of attribute indices and if the second option is 'false', then this is the list of attributes removed. If 'InvertSelection' option is changed to 'True' the list of indices are the attributes that are not removed.

Write '3' into the field 'attributeIndices' and click on the **OK** button. The window with the filters options closes. Now click on the **Apply** button on the right. The filter now removes the attribute with index three from the loaded data set. This change does not affect the data set in the file. The changed data set can be saved to a new ARFF file by pressing the **Save...** button and entering a file name. The action of the filter can be undone by pressing the **Undo** button.

Instead of calling the 'Remove' filter, attributes can be removed by selecting them in the small box in the 'Attributes' field and pressing the **Remove** button below the list of attributes.

5.4.1 Exercises

Ex. 6: Load the 'weather.nominal' data set.

Select the filter 'weka.unsupervised.instance.RemoveWithValues' and use it to remove all instances where the attribute 'humidity' has the value 'high'. (Click on the field beside the 'Choose' button, that is the field with the text 'RemoveWithValues' in it and change the attributes of the filter.)

Ex. 7: Undo the change to the data set that was done in exercise 6.

6 The Visualize Panel

Load the data set 'iris.arff' The iris data is based on flower measurements. Each instance is classified as one of the three types—iris-setosa, iris-versicolor or iris-virginica. The 'iris' data has 150 instances (50 examples of each type of iris).

Each instance has 5 attributes: 4 numeric attributes describing the dimensions of the flower, and the class attribute.

Click on the **Visualize** tab to change to the Visualizer panel. The Visualizer first shows a matrix of 2D-pixel plots with every possible combination of the five attributes of the 'iris' data on the x and y-axis. Click on the first plot in the second row and a windows will open, showing an enlarged 2D scatter plot using the selected axes. The instances are shown as little crosses. The colour of the cross depends on the class of the instance. The value at the x-axis is the 'sepallength' attribute and the value at the y-axis is the 'petalwidth' attribute.

Clicking on one of the crosses opens a 'Instance Info' window. In this window for the selected instance the values of all attributes are listed. **Important: Unfortunately, the instance number given in the 'Instance Info' window is different (-1) to the instance number in the 'Viewer'** Close the 'Instance Info' window again.

On the top of the window the select fields can be used to change the attributes on x-axis and y-axis. The left field in the second row allows you to change the coding of the colour. Try to change the x-axis to 'petalwidth' and the y-axis to 'petallength'.

Each of the little plots at the right represent a single attribute. Right clicking on one selects that attribute for the y-axis of the scatter plot. Left clicking on one does the same for the x-axis. Try to change the x- and y-axis back to 'sepallength' and 'petalwidth' using these plots.

The jitter bar moves the cross for each instance randomly from its position and can reveal where instances lie on top of each other. Experiment a little by moving the bar.

The Select Instance button and the Reset, Clear and Save buttons let you change the data set. Instances can be selected and removed. Try with selecting 'Rectangle' with the select instances button and select an area by left-clicking and dragging the mouse. Release as soon as the rectangle is the size that it covers the instances you want to select. The Reset button changes into a Submit button. Click on 'Submit' and all instances outside of the rectangle are deleted. Use 'Save' to save the data set to file. 'Reset' to restore the whole data set.

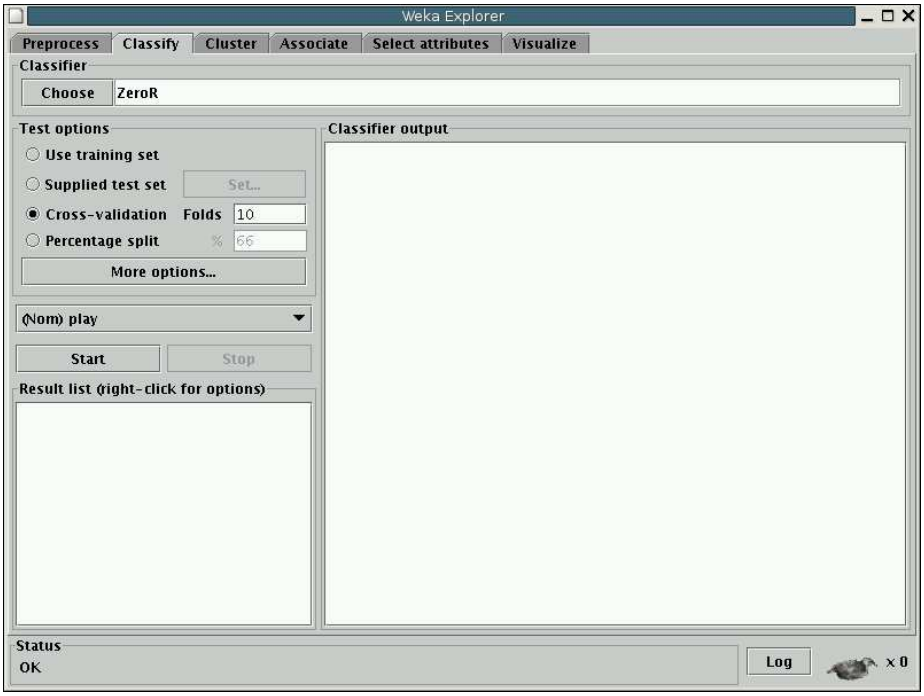
7 The Classify Panel

Now you know how to load a data set from a file and how to visualize it in 2D-plots. Next we will apply a machine learning algorithm (called a classifier) to the data. The classifier builds (learns !) a model from the data. Classifiers help us to classify data automatically.

Before getting started we need to load the 'weather.nominal' data set again ('Open file' button, select 'weather.nominal.arff').

To do classification tasks we have to change to a different panel. Click on the **Classify** tab at the top of the window to switch to the classification panel

of the WEKA Explorer.



A very popular data mining (machine learning) method that builds decision trees is 'J48'. Choose the 'J48 classifier' by clicking on the **Choose** button on the top of the window. A dialog window will appear with various types of classifiers to choose from. Click on the circle left beside the **trees** entry in the list and the sub-entries will appear. Click on **J48** to choose it. (Classifier names, like filter names are organized in a hierarchy. 'J48' has the full name 'weka.classifiers.trees.J48'.)

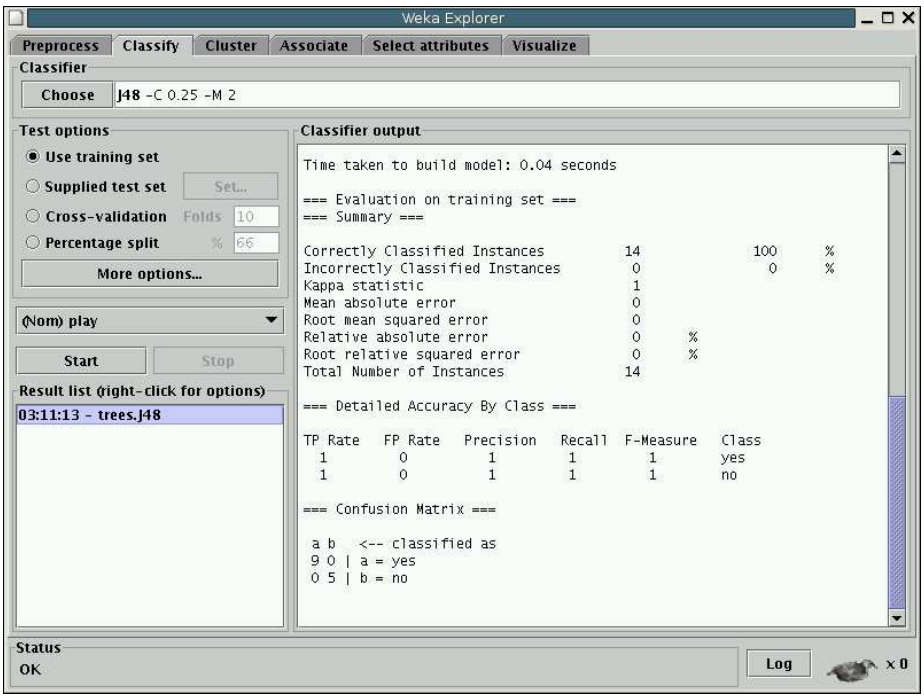
The name of the classifier listed in the text box right beside the 'Choose' button should now read 'J48 -C 0.25 -M 2'. The text after 'J48' represents parameter settings of the J48 classifier. We can ignore these, since for our first test the defaults are sufficient.

Decision trees are a special type of classification model. This model should be able to predict the class of new instances given to it with high accuracy. Accuracy is the percentage of correctly classified instances. After a model has been learned, we should test it to see how accurate it is on new data. For this we will need a test set. The parameters of how to test are set in the **Test options** panel. First we want to choose the 'Use training set' option. For that click on the round button left of 'Use training set'.

What is meant by training set and test set? To build a model we need a training data set, and in the WEKA Explorer the data set that is loaded in the 'Preprocess' panel ('weather.nominal.arff' at this time) is taken as training set. It is called the training set because building a model can also be called 'training' a model. After the training the testing is done. For each instance of the test set

the model is used to predict a class value, although this instances already have a class value. Summing up the cases where the predicted value was the same as the real value gives the accuracy measurement.

We have set the test options so that we use the same data set for testing as for training. We are now ready to train our first classifier. Do so by clicking on the **Start** button.



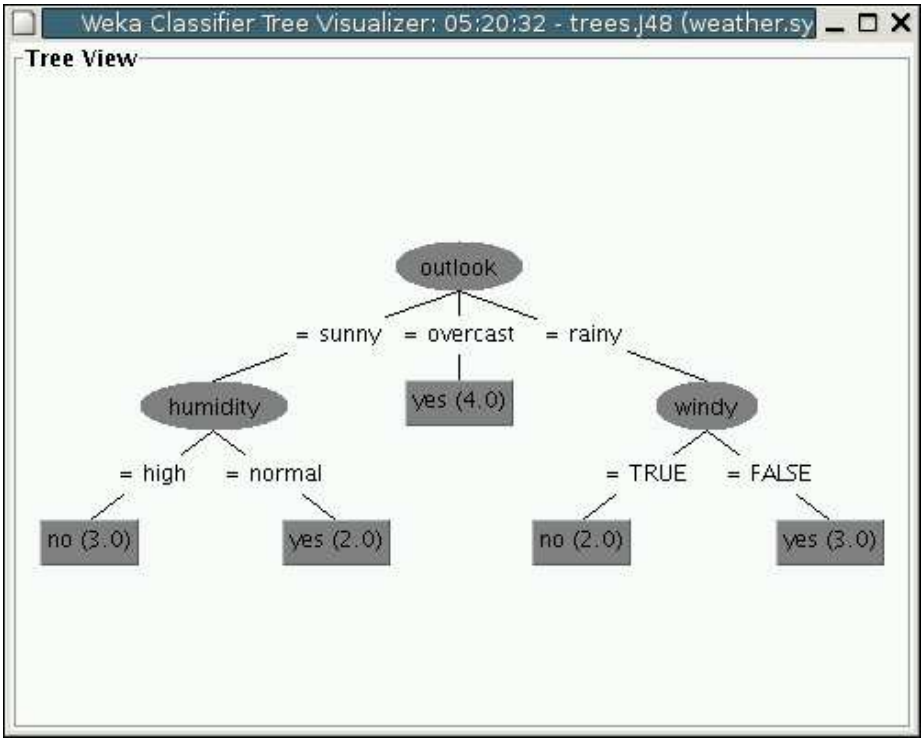
The results of the training and testing of this model will appear as text in the **Classifier output** box on the right. You can scroll through the text to examine it. We want to consider only three parts of the output. First let us look at the part that describes the decision tree that was generated:

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

outlook = sunny
|   humidity = high: no (3.0)
|   humidity = normal: yes (2.0)
outlook = overcast: yes (4.0)
outlook = rainy
|   windy = TRUE: no (2.0)
|   windy = FALSE: yes (3.0)

Number of Leaves  :  5
```

```
Size of the tree :  8
```

This text represents the decision tree that was built, but this representation is not very easy to interpret. Let us look at a graphical representation of this tree. Each time we press the 'Start' button the classifier is trained and tested again and a new entry is written into the **Result List** panel in the lower left corner.

Right-click on the **trees.J48** entry in the result list and choose **Visualize tree**. A window pops up that shows the decision tree as a graph. Right-click a blank space in this window to bring up a new menu enabling you to auto-scale the view, or force the tree to fit into view. Dragging the mouse lets you pan around.



The instances are classified using this tree. The first condition is the one in the so called root node at the top. 'outlook' is the attribute tested and, depending on the value of this attribute, the testing continues down one of the 3 branches. If the value was 'overcast' the testing ends and the predicted class is 'yes'. The last nodes are the leaf nodes and they give the class that is to be predicted. Let us go back to the root node. With the value 'sunny' the attribute 'humidity' is tested, and with the value 'rainy' the attribute 'windy' is tested. None of the paths through the tree have more than two tests so the decision tree is quite simple.

Now back to the 'Classifier output' window. The next two parts of the

output give a report on what the testing found out about the quality of this model.

The following states how many test instances have been correctly classified. This is the accuracy of the model on the test set. It is 100%, which is often the case if the training set and test set are the same.

```
Correctly Classified Instances          14               100%
```

At the bottom of the output is the confusion matrix:

```
=== Confusion Matrix ===

 a b   <-- classified as
 9 0 | a = yes
 0 5 | b = no
```

Each element in the matrix is a count of instances. Rows in the matrix represent the true class of the instances, and columns represent the predicted class. As you can see, all 9 'yes' instances have been predicted as yes and all 5 'no' instances as no.

7.0.2 Exercise(s)

Ex 8: How would the following instance be classified using the given decision tree?

outlook = sunny, temperature = cool, humidity = high, windy = TRUE

7.1 Setting the Testing Method

The 'Test options' box gives several possibilities for testing a data set. You have to know them for the following tutorials. After the **Start** button is pressed the classification method selected is started and the data set that was loaded in the 'Preprocess' panel is used for training a model. It then uses a test set for testing the model. The 'test options' box offers several testing methods.

Use training set Uses the same data set that was used for training (the one that was loaded in the 'Preprocess' panel) again as test data set.

Supplied test set Let you enter a data set to use as test set.

Cross-validation Splits the training set into folds (disjoint subsets). The number of folds can be entered in the Folds field. It takes all but one of the folds for training and the left out one for testing. It then takes a new fold for testing and the rest for training and repeats until all folds have been used for testing exactly once.

Percentage split Splits the training set and takes the percentage entered as training data set.

All testing methods except the cross-validation method perform classification only once. The cross-validation method performs classification 'fold' number of times and gives as results the average of the 'fold' results.

7.1.1 Exercise(s)

Ex 9: Change the test set: In the **Test options box** choose the **Supplied test set** option, and click on the **Set...** button.

A small window will appear for choosing the test set. Click **Open file...** and browse to open the file named 'weather.test1.arff'. Click **Open** to select the file. You can close the small window to return to the main WEKA window.

This test file contains the 3 instances you see below. Press the 'Start' button to train and test again. Did the decision tree change? How many instances are correctly classified? Interpret the confusion matrix.

Relation: weather.symbolic

No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	**play** Nominal
1	sunny	cool	high	TRUE	yes
2	overcast	mild	high	FALSE	yes
3	rainy	cool	high	FALSE	yes

7.2 Visualizing Classification Errors

Classification errors can be viewed in the Visualizer. Right-click on the **trees.J48** entry in the result list and choose **Visualize classifier errors**. A visualizer window pops up. The instances are represented as crosses and if incorrectly classified as little squares. Since all attributes are nominal and each attribute has only a few values, in most plots instances are printed on top of each other. Use the 'Jitter'-bar to see how many are on one spot.

7.2.1 Exercise(s)

Ex 10: Use the 'Visualize classifier errors' functions to find the wrongly classified test instance of Exercise '9'. What is the instance number of this instance? **(Remember, the instance number given in the 'Instance Info' window is different (-1) to the instance number in the 'Viewer')**

8 Answers To Exercises

1. Hot, mild and cool.

2. The iris data set has 150 instances and 5 attributes. So far we have only seen **nominal** values but the attribute petallength is a **numeric** attribute and can have any real value. In this data set the values for this attribute are between 1.0 and 6.9 (see in the right panel 'Minimum' and 'Maximum').

3. The class value of this instance is 'no'. (If you start the data set editor pressing the **Edit** button the **Viewer** window opens. In this window the instances are listed. The first column is the instance number. The row with the number 8 in the first column is the instance with instance number 8.)

4. It is not an attribute of the data set. 'No.' is the instance number that is given to an instance after it has been loaded from the arff file. It corresponds with the order the instances are in the ARFF-file.

5. This can be easily seen in the **Viewer** window. The iris data set has four numeric and one nominal attributes. The nominal attribute is the class attribute.

6. Load the 'RemoveWithValues' filter after clicking the **Choose** button. Click on the field right to the 'Choose' button and set the field 'attributeIndex' to 3 and the field 'nominalIndices' to 1'. Press 'Ok' and 'Apply'.

7. Click the 'Undo' button.

8. The test instance would be classified as 'no'.

9. The training set has not been changed so the decision tree did not change either. The number of correctly classified instances is 2. One instance of class 'yes' has been classified as 'no'.

10. Instance No. 0 (0 in the scatter plot but 1 in the Viewer of the Preprocess panel) is wrongly classified and is therefore represented as a square in the 2D scatter-plot.