《Data mining》Lab 1

Weka and decision tree classifier

Class **计科**11606  Name \_\_**张仁兵**\_\_ Score \_\_\_\_\_\_

1.0 Background reading

[Weka machine learning workbench](http://www.cs.waikato.ac.nz/~ml/weka/index.html)

1.1 Learning goals

(1) Install Weka and learn its interface.

(2) Load the weather dataset and look at its attributes.

(3) Build a decision tree classifier, visualize it and use it to classify new instances.

1.2 Download and install Weka

* Download Weka from the [download page of Weka](http://www.cs.waikato.ac.nz/~ml/weka/downloading.html). Choose an appropriate stable version.
* Run Weka 3.8 version (32 bit or 64 bit, with or without Java VM) installer. Choose default options.
* Go to Program Files folder and find Weka. Create a shortcut for Weka/data to desktop. This folder contains the datasets that come with Weka.

1.3 The Explorer interface

* After you have started up the WEKA Explorer you can see that the user interface contains six panels. At the top of the window are the tabs of these panels and the Preprocess panel is the one that is open when the Explorer has first been started. This tutorial will introduce you to three of these six panels: the Preprocess panel, the Classify panel, and the Visualize panel. The other three panels will be covered in later tutorials. What follows is a brief description of the functions that the three panels perform.
* **Preprocess** Enables you to load and pre-process data. After a dataset has been loaded, the panel displays certain information about it. The data can also be modified, either by editing it in the dataset editor or by applying a filter. The modified dataset can also be saved. As an alternative to loading a pre-existing dataset, an artificial one can be created by using a generator. It is also possible to load data from a URL or a database.
* **Classify** From this panel the classification methods in WEKA can be started. Several options for the classification process can be set and the results of the classification can be viewed. The training dataset used for classification is the one loaded (or generated) in the Preprocess panel.
* **Visualize** The dataset loaded in the Preprocess panel can be visualized in two-dimensional scatter plots in this panel. The user selects the attributes for the x-axis and the y-axis of the plot.

1.4 The ARFF format

* Before changing to any of the other panels, the Explorer must have a dataset to work with. To load up a dataset, click on the Open file... button in the top left corner of the panel. Inside the C:\Program Files\Weka-3-6\data folder you will find a file named weather.nominal.arff, which contains the nominal version of the “weather” dataset discussed in class. Open this file.
* The weather data is a small dataset with only 14 examples for learning. Because each row is an independent example for learning, the rows/examples are called “instances”. The instances of the weather dataset 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 sub panel the panels on the right will show information about the selected attribute. You can see the values of the attribute and how many times an instance in the dataset has a particular value. This information is also shown in a histogram.
* All attributes in this dataset are nominal, i.e. they have a pre-defined set of 2unordered values. The data in one instance describes a weather forecast for a particular day and whether to play a certain game on that day. It is not really clear what that game is, but let us assume that it is golf. 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.
* **What are the values that the attribute ‘temperature’ can have?**

**Three value :Hot 、mild、cool**

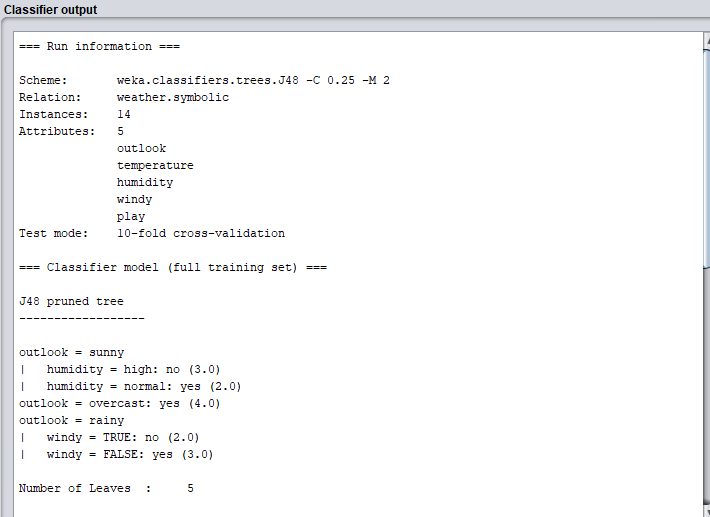
* Load a new dataset: Press the ‘Open file’ button and select the file C:\Program Files\Weka-3-6\data\iris.arff. How many instances does this dataset have? How many attributes does it have? What is the range of possible values of the attribute ’petallength’?

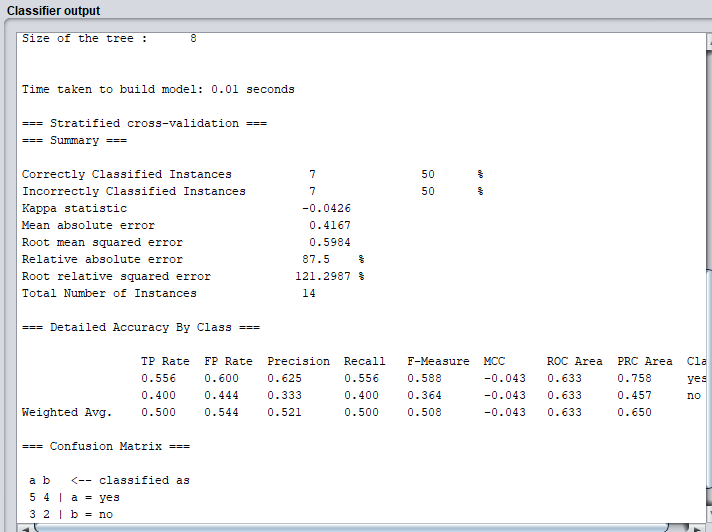
**Irirs’ dataset hava 150 instances,and hava 5 Attributes;**

**The attribute of Petallength has four values.**

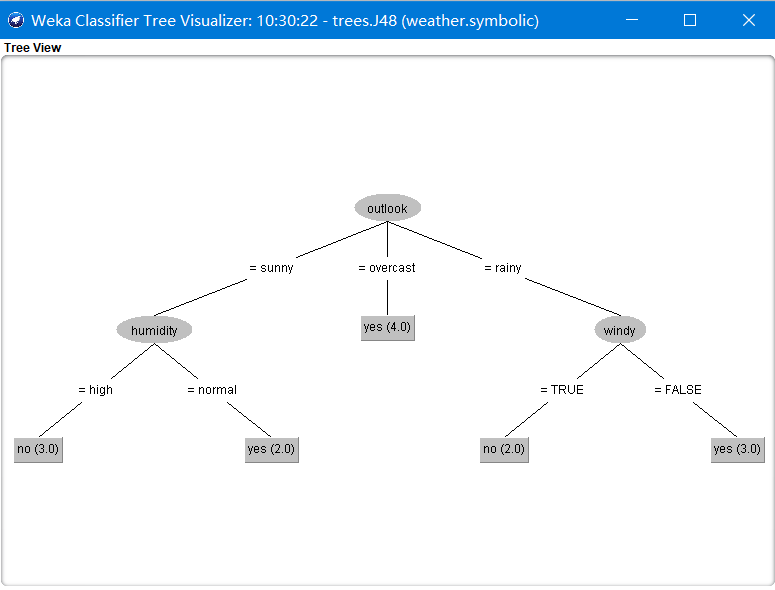
1.5 Build a decision tree classifier

* Now you know how to load a dataset from a file. In this section we will apply a classification algorithm to the data—called “classifier” in WEKA. The classifier builds (“learns”) a classification model from the data. Note that all schemes for predicting the value of a single attribute based on the values of some other attributes are called classifiers in WEKA, even if they are used to predict a numeric target (i.e. even if they would perhaps be more accurately described as numeric prediction or regression schemes). The reason is that, historically and in the context of machine learning, numeric prediction has also been called “classification with continuous classes”.
* Before getting started, we need to load the weather data again. Go to the Preprocess panel, click the Open file button, and select weather.nominal.arff from C:\Program Files\Weka-3-6\data.
* To perform 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 machine learning method for data mining, one that builds decision trees, is the C4.5 algorithm. It is implemented in a classifier called “J48” in WEKA. Choose the J48 classifier by clicking on the Choose button close to the top of the Classifier tab. A dialogue window will appear with various types of classifiers to choose from. Expand the trees entry in the list and sub-entries will appear. Click on J48 to choose the J48 classifier. Note that classifier names, like filter names, are organized in a hierarchy: J48 has the full name weka.classifiers.trees.J48.
* The name of the classifier is shown in the text box next to the Choose button. It should now read J48 -C 0.25 -M 2. The text after J48 represents (default) parameter settings of the J48 classifier. We can ignore these, since the defaults rarely require changing to obtain good performance with C4.5.
* Decision trees are a special type of classification model. Ideally, the model should be able to predict the class values of new, previously unseen instances with high accuracy. In classification tasks, accuracy is often measured as the percentage of correctly classified instances. After a model has been learned, we should test it to see how accurate it is when classifying instances.
* One option in WEKA is to evaluate on the training set, the data that is used to build the classifier. This is NOT generally a good option because it will lead to optimistic performance estimates: we can get 100% accuracy on the training data by simple rote learning, but this does not tell us anything about performance on new data that we might encounter when we apply the model to make predictions in the future. Nevertheless it is instructive to consider predictive performance on the training data.
* In WEKA, the data that is loaded in the Preprocess panel is the “training data”. To evaluate on the training set, choose Use training set from the Test options panel in the Classify panel. Once the test strategy has been set, we can build the classifier and evaluate it by pressing the Start button. This will process the training set using the currently selected learning algorithm (i.e. C4.5). Then, it will classify all the instances in the training data, because this is the evaluation option we have chosen, and output performance statistics.
* **What is the output of this process?**

****

****

* This text represents the decision tree that was built, including the number of instances in each of the leaves. However, this representation is not very easy to interpret. Fortunately, we can generate a graphical representation of this tree. Note that each time we press the Start button and a new classifier is built and evaluated, a new entry is written into the Result List panel in the lower left corner. To visualize the tree graphically, we can use the entry that has just been added to the list. We can 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.
* **Take a screenshot of the visualized tree.**

****

* **How would the following instance be classified using the decision tree?**

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

