May 1, 2011

Data Mining A Tutorial-Based Primer

Chapter Four using WEKA

Most of the datasets described in the text have been converted to the format required by WEKA. These datasets can be found in the .zip file titled *iDA files formatted for Weka*. Also, the installed WEKA software includes a folder containing datasets formatted for use with WEKA. This folder contains ten datasets and is likely located in *c:\program files\weka-3-6\data.* Some of these datasets are used in the exercises. Finally, additional datasets formatted for use with WEKA can be downloaded at the Web site:

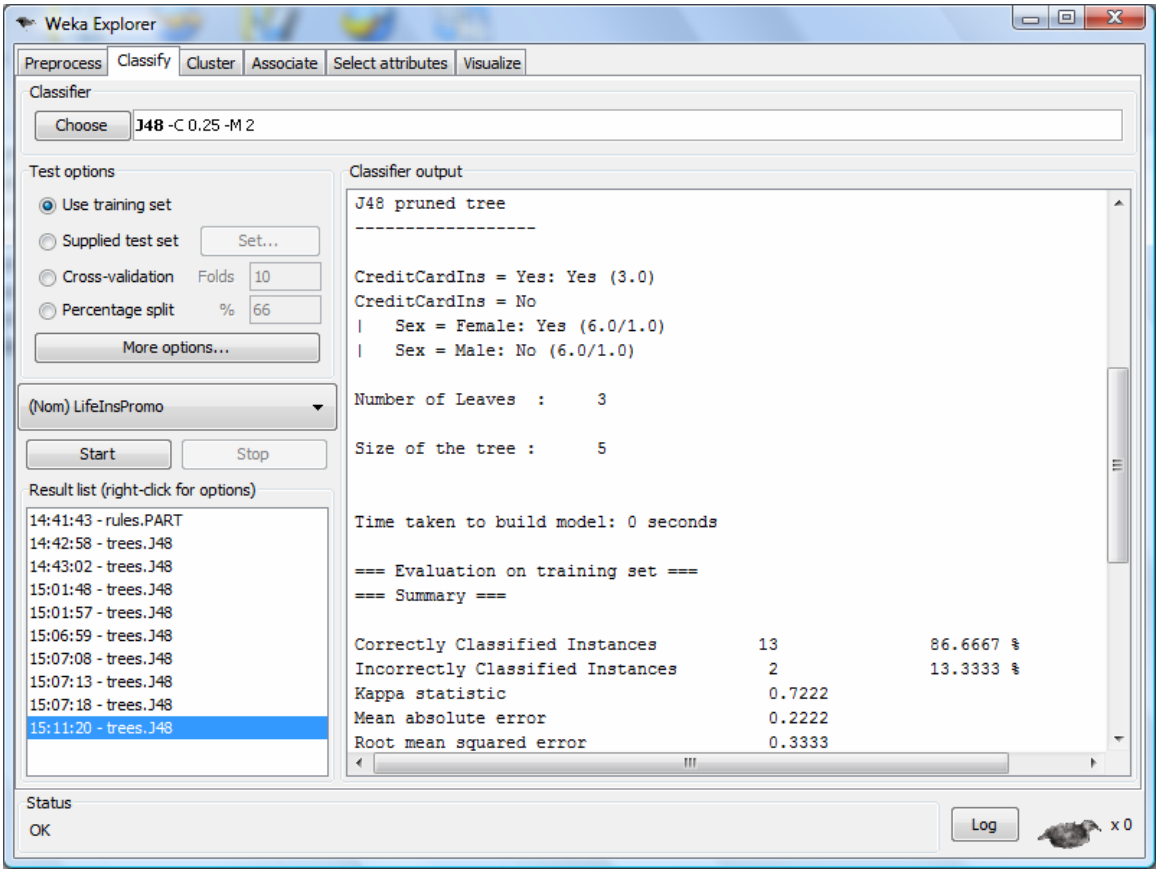
http://www.cs.waikato.ac.nz/~ml/weka/index.html

Here is a suggested methodology for incorporating WEKA into Chapter 4 of the text.

- Section 4.1:
    - Use Figure 4.1 to provide a general discussion of the components of iDA.
- Section 4.2:
    - The concept hierarchy is a common data structure for data mining. With the help of Figure 4.3, offer an overview of concept hierarchies and how they are used.
- Section 4.3: Skip.
- Section 4.4:
    - First, cover the following categorical data concepts listed in section 4.4 of the text:
        - Use Figure 4.8 to cover domain predictability.
        - Use Figure 4.10 to cover class predictability.
        - Use Figure 4.10 to cover class predictiveness.
        - Use page 124 to cover necessary and sufficient attributes.

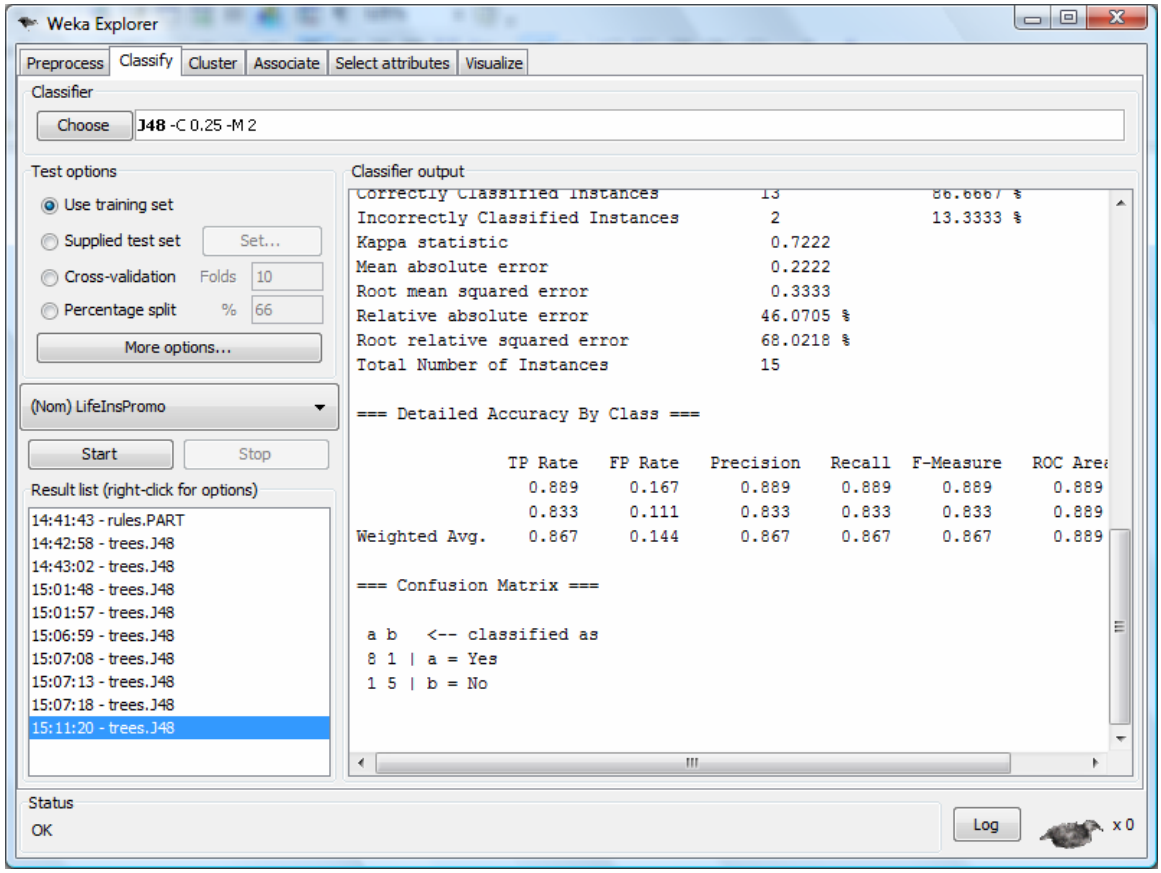
- Cover section 4.5 below prior to finishing the remainder of the material in section 4.4. Here's why:
    - WEKA's decision tree and rule based classifiers are easy to use and understand. Students will have a much better data mining experience if section 4.5 is covered prior to examining WEKA's unsupervised clustering tools.
- Once section 4.5 has been completed, replace the section 4.4 coverage of unsupervised clustering with a discussion of WEKA's K-Means algorithm ─ SimpleKMeans. A good way to explain unsupervised clustering with WEKA is to work through data mining exercise 6 in class.
    - As an option, Expectation Maximization (EM) can also be covered. EM is a more interesting unsupervised clustering algorithm and is described in the text on pages 315 through 317.

- Section 4.5:
    - Replace the coverage of supervised learning using ESX with a discussion of J48 ─ WEKA's decision tree building tool. J48 is similar to the C4.5 decision tree model described in section 3.1 of the text. J48 was used to generate the decision trees given in Chapter 2 of the text.
    - Let's look at an example using J48 with the ccpromo.arff dataset.
        - Initiate WEKA's Explorer and load the ccpromo.arff dataset.
        - Left click on Classify at the top of the screen.
        - Left click on Choose and select J48 listed under the Trees folder.
        - Set *Life Insurance Promo* as the class (output) attribute. Set *Test options* to *Use training set.*
        - Click on *Start.* The created decision tree appears as below.
    -



o

- Notice that *CreditCardIns* is the attribute chosen for the Top-level node of the tree. It takes a few data mining sessions to get used to WEKA's output format for decision tree structures.
- Compare the decision tree above to the tree given on page 75 of the text ─ they are the identical.
- Thirteen of the fifteen instances were correctly classified.
- Continue to scroll until your screen appears as below.

- 

    - The confusion matrix tells us that one individual accepting the life insurance promotion was incorrectly classified as a reject and one individual rejecting the promotion was incorrectly classified as accept.

    - Next, try the same experiment as above but Set *Test options* to *cross-validation.* Try several experiments by changing the *Folds* value.

- Section 4.6:

    - Replace the coverage of Rulemaker with a discussion of PART ─ WEKA's rule generator.

    - An example using PART is given in the WEKA tutorial document.

- Section 4.7:

    - Use Figure 4.13 to provide a general discussion of instance typicality.

- Section 4.8: Skip.

Chapter Four Exercises - WEKA

Review Questions

1. Differentiate between the following terms.

    a. Class predictiveness and class predictability
    b. Domain predictability and class predictability
    c. Within-class and between-class measure

2. Suppose you have used data mining to build a model able to differentiate between individuals likely and unlikely to default on a car loan. For each of the following, describe a categorical attribute value likely to display the stated characteristic.

    a. A categorical attribute value that is necessary but not sufficient for class membership.
    b. A categorical attribute value that is sufficient but not necessary for class membership.
    c. A categorical attribute that is both necessary and sufficient for class membership.

Data Mining Questions

1. You suspect marked differences in promotional purchasing trends between female and male Acme credit card customers. You wish to confirm or refute our suspicion. Perform a supervised data mining session using the CreditCardPromotion database in conjunction with PART. Use *sex* as the output attribute. Designate all other attributes as input attributes, and use all 15 instances for training. Write a summary confirming or refuting our hypothesis. Base the analysis on rules created for each class.

2. Repeat the previous exercise using J48 rather than PART but base the analysis on the created decision tree.

3. For this exercise you will use WEKA's J48 decision tree algorithm to perform a data mining session with the cardiology patient data described in Chapter 2. Open the WEKA explorer and load the cardiology-weka.arff file. This is the mixed form of the dataset containing both categorical and numeric data. Recall that the data contains 303 instances representing patients who have a heart condition (sick) as well as those who do not.

Preprocess Mode Questions:

a. How many of the instances are classified as Healthy?
b. What percent of the data is female?
c. What is the most commonly occurring domain value for the attribute *slope*?
d. What is the mean age within the dataset?
e. How many instances have the value 2 for # of Colored Vessels?

Classification Questions using J48:

Perform a supervised mining session using 10 fold cross validation with J48 and *class* as the output attribute. Answer the following based on your results:

a. What attribute did J48 choose as the top-level decision tree node?
b. Draw a diagram showing the attributes and values for the first two levels of the J48 created decision tree.
c. What percent of the instances where correctly classified?
d. How many healthy class instances were correctly classified?
e. How many sick class instances were falsely classified as healthy individuals?
f. Determine how True Positive Rate (TP Rate) and False Positive Rate (FP Rate) are computed.

Classification Questions using PART:

a. List one rule for the healthy class that covers at least 50 instances.
b. List one rule for the sick class that covers at least 50 instances.
c. List one rule that is likely to show an inaccuracy rate of at least 0.05.
d. What percent of the instances where correctly classified?
e. How many healthy class instances were correctly classified?
f. How many sick class instances were falsely classified as healthy individuals?

4. Load the CreditScreening dataset described on page 163 of the text into the WEKA Explorer. Make sure that *class* is designated as the output attribute. Use J48 together with 10-fold cross validation to mine the data. Record your results including the attributes used to create the root node and first level of the decision tree. Mine the data a second time but this time use only the root attibute and the first level attributes used in the first mining session ─be sure to also keep the *class* attribute. Write a short summary comparing your results. Can you draw any general conclusions from your experiment?

5. Use Wordpad or MS Word to open the *soybean* dataset located in the folder ─ *c:\program files\weka-3-6\data.* This dataset represents one of the more famous data mining successes. Classification accuracy of unseen instances is likely to be above 90% with most classifiers.

   Scroll through the file to get a better understanding of the dataset. Open WEKA's Exporer and load this dataset. Classify the data by applying J48 with a 10-fold cross validation. Report your results.

6. For this exercise, you will use WEKA's SimpleKMeans unsupervised clustering algorithm with the heart disease dataset.
   a. Open the WEKA Explorer and load the numerical form of the heart disease dataset ─ CardiologyN-Weka.
   b. Remove the Class attribute as you do not want the value of this attribute to affect the clustering.
   c. Click on *Cluster* and choose the *SimpleKMeans* algorithm
   d. Invoke the *object editor* with a left click in the white space area of the *choose* bar.
   e. Set *displayStdDevs* to *True.* This will give us the domain standard deviation of each attribute as well as the within-class attribute standard deviations.
   f. We know there should be two distinct clusters, set numClusters to 2.
   g. Click on *Start* to begin the data mining session.
   h. Copy the results into a word document. The output should include attribute mean and standard deviation values for each cluster as well as the total number of instances assigned to each cluster.
   i. Next, you must decide if the clusters are interesting. You can use a rough measure of attribute significance to accomplish this. Specifically, for each attribute, subtract the attribute means for the two clusters and divide the absolute value of this result by the domain standard deviation for the attribute. For example, the computation for attribute *age* will likely be | 56.1216 – 54.3663 | / 9.0921 = 0.19. Computations near or greater than one indicate attributes that have been clearly differentiated by the clustering. If there are no such attributes, the clustering is of little interest. After making the computations, make a general statement about whether the clustering merits further exploration.

7. Repeat exercise 6 using WEKA's Expectation Maximization (EM) algorithm. EM's output does not include domain attribute standard deviations. Therefore, to perform the analysis given in exercise 6, you will need to first work through exercise 6. As an alternative, those that have some familiarity with statistics can use the method given on page 232 of the text.

Computational Questions

1. Concept class C_1 shows the following information for the categorical attribute *color*. Use this information and the information in the table to answer the following questions:

Name	Value	Frequency	Predictability	Predictiveness
Color	Red	30		0.4
	Black	20		1.0

   a. What percent of the instances in class *C* have a value of *black* for the *color* attribute?
   b. Suppose that exactly one other concept class, C_2, exists. In addition, assume all domain instances have a color value of either *red* or *black*. Given the information in the table, can you determine the predictabilty score of *color = red* for class C_2? If your answer is yes, what is the predictability value?
   c. Using the same assumption as in part b, can you determine the predictiveness score for *color = red* for class C_2? If your answer is yes, what is the predictiveness score?
   d. Once again, use the assumption stated in part b. How many instances reside in class C_2?