Department of Computer Engineering
University of Peradeniya

CO544 Machine Learning and Data Mining

Lab 4: Classification and prediction using WEKA

December 6, 2018

Part 1: Classification

- **Aim:** This part of the lab aims to provide students hand on experience in classification using WEKA data mining toolkit.
- **Objectives:** At the end of this part of the lab students should be able to,
  - Perform simple preprocessing tasks using filters available in WEKA.
  - Use various classification algorithms with different test options for a given classification problem.
  - Analyze the output of the classification algorithm and interpret the results.

1. Load the **Zoo** data set. Observe attributes and their values.
2. Build the **C4.5** decision tree using default parameters and test options. Observe the output of the algorithm.
3. Visualize the output of **C4.5** algorithm. Explore different error estimates and record the classification accuracy of C4.5 algorithm. Examine the true positive (TP) rates, the false positive (FP) rates and the confusion matrix. Explain misclassifications observed in the confusion matrix.
4. Evaluate the **C4.5** algorithm using the following testing options:
   (a) The training set,
   (b) 10-fold cross validation.

   Record the classification accuracies using both methods. Which one provides more realistic future performance? Why?
5. Can you apply the **ID3** learning algorithm on this data set? Explain your answer.
6. Remove **animal name** and **legs** attributes from the data set using 'Remove' filter available in **preprocess** tab. For that, select attribute indices and click on **'Remove'** button to remove from the data set. You can save the modified data set using **'Save'** button.

7. Now build the **ID3** decision tree. Examine the output. Record the 10-fold cross-validation accuracy.

8. Use **OneR** algorithm and explain the classifier output. Record the 10-fold cross validation accuracy.

9. Apply the **ZeroR** algorithm to the **Zoo** data set. Observe the output and then explain the model generated by the **ZeroR** algorithm. In order to find more information on the algorithm click on the algorithm name beside the **'Choose'** button in classifier and then click on the **'More'** button.

10. Use another classification algorithm of your choice and observe the output of the algorithm. Compare the results of the chosen algorithm with previous outputs.

Part 2: Predicting Class Values

- Aim: This part of the lab aims to provide students hand on experience in using WEKA data mining toolkit to predict class labels for a given data set.

- Objectives: At the end of this part of the lab, students should be able to
  - Develop a classification data model using a training data set and make predictions for a testing data set.
  - Analyze the output and interpret the results.

In WEKA tool, after a model has been learned under the classification learning, one can make predictions for a test set. If the class values are present in a test set the output will contain both the actual and predicted class values. If the class values are missing from a test set, the actual class label for each instance will not contain useful information, but the predicted class label will. We will demonstrate how to do class prediction using WEKA for a test set (which includes the class values) using **`zoo_train.arff`** and **`zoo_test.arff`** files.

The steps are as follows.

1. Load the **`zoo_train.arff`** data set. Observe the attributes and their values.

2. Build the **C4.5** decision tree (J48 in WEKA) with **'Use training set'** test option.

3. Select **'Supplied test set'** test option. Click on the **'Set'** button to the right side of that option. Then a separate window named **'Test Instances'** will pop out. Click on **'Open file'** button and browse and select the **`zoo_test.arff`** file. Then close the **'Test Instances window'**.

4. Click on **'More options'** button. Then a separate window named **'Classifier evaluation options'** will pop out. Select **'Output predictions'** and choose **'PlanText'** and click **'OK'**. When you click **'Start'**, Classifier output contains Predictions on test set as in Figure 1.

   Note: If you want to re-evaluate the model on new test set, right click on the result buffer in the **'Result list'**, which corresponds to the model and select the option **'Re-evaluate model on current test set'** option from the drop down list. After this, the classifier output will contain the predictions for the new test instances.

```
=== Predictions on test set ===

    inst#     actual  predicted error prediction
        1 7:invertebrate 7:invertebrate       0.875
        2     4:fish     4:fish       1
        3     2:bird     2:bird       1
        4   1:mammal   1:mammal       1
        5 7:invertebrate 7:invertebrate       0.875
        6     4:fish     4:fish       1
        7     2:bird     2:bird       1
        8   6:insect 7:invertebrate   +   0.875
        9 5:amphibian 5:amphibian       1
       10  3:reptile 7:invertebrate   +   0.875
       11  3:reptile 5:amphibian   +   1
       12     4:fish     4:fish       1
       13   1:mammal   1:mammal       1
       14   1:mammal   1:mammal       1
       15     2:bird     2:bird       1
       16   1:mammal   1:mammal       1
       17   6:insect   6:insect       1
       18   1:mammal   1:mammal       1
       19 7:invertebrate 7:invertebrate       0.875
       20     2:bird     2:bird       1
```

Figure 1: Predictions on Zoo test data set

You can interpret the above results as follows.

- First column - instance numbers.
- Second column - actual class value of each test instance.
  - If class is not present, a '?' symbol will be displayed.
  - If class is present the class number with its class value will be displayed. For example in the first instance actual class value is displayed as '7:invertebrate'. This means instance one is predicted to be of class 7, whose value is invertebrate.
- Third column - predicted class value of each test instance.
- Fourth column - whether the predicted class value mismatches with the actual class value.
  - If it is an error a '+' sign will be displayed. Else nothing will be displayed.
- Fifth column - estimation of probabilities for each instance actually belongs to a class.

Note: Here we first built the model using the train set and used a test set for predictions at the same time. If you want, you can build the model and save it for future uses. Then you may load it at a later time for predictions.

Exercise

Find class labels for the data set given in **`zoo_test_classmissing.arff`**. This file has the same data set with missing class values. Then compile a report as a text document with answers to questions raised in the lab tasks. Name the report as **`e14xxxlab4.txt`** where xxx refers your registration number.