Lab Exercise (Weka)

CIS400/600 Fundamentals of Data and Knowledge

Mining

Problem 1 - Data Preprocessing

1. Download and install **Weka** from `http://www.cs.waikato.ac.nz/ml/weka/downloading.html`
2. Download the classic iris dataset (iris.csv) from `https://raw.githubusercontent.com/uiuc-cse/data-fa14/gh-pages/data/iris.csv`
3. Weka uses **.arff** file format. Convert the above downloaded **iris.csv** to **iris.arff** using Weka. Submit the file **iris.arff**

Problem 2 - Decision Tree Construction and Visualization

1. Write the Weka command to construct decision tree using J48 with 10-fold cross validation
2. How many misclassified instances exist when 10-fold cross validation is used
3. What is the percentage of *Accuracy*
4. Visualize the tree and specify depth of tree, no. of leaf nodes
5. How many instances are classified under *setosa, virginica, versicolor*
6. For petal_width <= 0.6, what is the class label under which the instances are classified

7. For petal_width <= 1.7, how many instances are classified under *versicolor*

8. For petal_width > 1.7, how many instances are classified under *virginica*

Problem 3 - Decision Tree Construction and Visualization

1. Choose the **breast-cancer** dataset from Weka viz. **breast-cancer.arff** and perform decision tree learning using **Random Forest** classifier

2. In **Classifier Output** on the right, it is mentioned that **Correctly Classified Instances** a.k.a *Accuracy* is **69.58%**. The fraction $\frac{x}{y} = 69.58\%$. What are the values of **x** and **y** in this case that led to 69.58%

3. Set the number of iterations options *numIterations* to 50, 100, 150, 200, 250. List *Accuracy* for each of these options. What is the trend that is observed in accuracy values.

4. Perform classification on breast-cancer dataset using **J48**. Which among **Random Forest** and **J48** has better accuracy

Problem 4 - Decision Tree Pruning

1. What is the Weka command to construct a tree by not pruning it

2. What is the Weka command command to prune a tree by allocating a minimum of 10 objects to a leaf node

3. What is percentage accuracy of decision tree if pruning is allowed and a minimum of 10 objects are to be allocated to a leaf node