**EIE3112 Database Systems**

**Lab 3: Data Mining using Weka on PolyU**

**Virtual Lab Platform - Lab Report**

**Name: Sze Kin Sang**

**Student No.: 19066206D**

**Objectives and Outcomes**

* After finishing this lab, you should be able to
* Use Weka to perform data mining
* Understand the differences between classification, cluster analysis, and regression analysis.
* Understand the characteristics of different classifiers and clustering algorithms

**Part I: Classification**

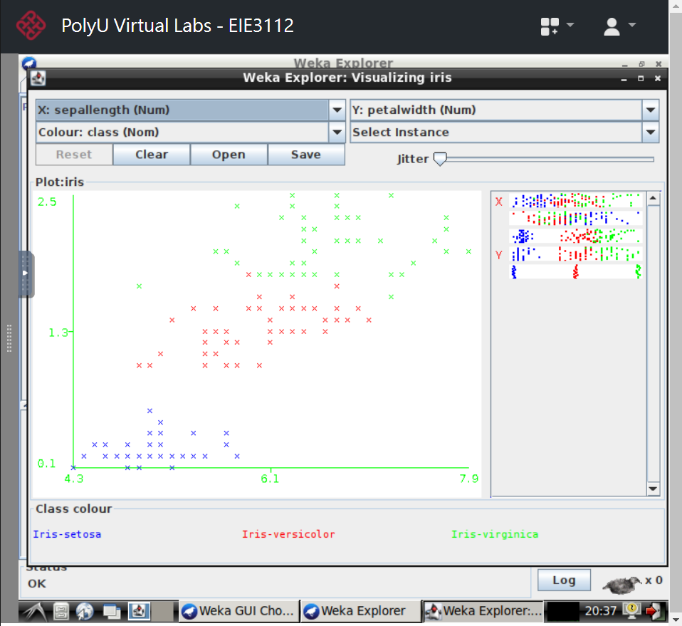
**Procedure:**

**1)**

**Chart

Description automatically generated**

Open the WEKA explorer and select the iris.arff file



Visualizing the data by selecting the suitable attributes

**Question 1: Assume that petalwidth and sepallength are used as features. Can linear classifiers such as linear support vector machines classify the training data perfectly? Explain your answer. Include screen capture for your explanation.**

According to the graph, it is easy to draw a decision plane to separate class Iris-setosa from other data. However, some of the data of class Iris-versicolor and class Iris-virginica overlap each other, which difficult to draw a decision plane. Thus, linear support vector machines cannot classify the training data perfectly.

Graphical user interface, application

Description automatically generated

**2)**

**Text

Description automatically generated with low confidence**

Train a decision tree by using J48 program

**Question 2: What is the accuracy (correctly classified instances) on the test data? Capture the portion of classifier output that contains the accuracy and circle it. Then put it in your Lab report.**

The accuracy on the test data is 96.0784%.

**Text

Description automatically generated with low confidence**

**3)**

**Diagram

Description automatically generated**

Visualizing the trained decision tree

**Question 3: Identify the attribute(s) that is (are) not relevant to the classification task. Explain your answer**

Sepallength and sepalwidth are not relevant to the classification task since they didn’t appear in the decision trees, which means the classes can be identify without these two attributes.

**4)**

**Text

Description automatically generated with low confidence**

Train an artificial neural network by using MultilayerPerceptron program

**Question 4: What is the accuracy of MultilayerPerceptron on the iris dataset? Capture the portion of classifier output that contains the accuracy and circle it. Then put it in your Lab report.**

The accuracy on the test data is 98.0392%.

**Text

Description automatically generated with low confidence**

**5)**

Graphical user interface

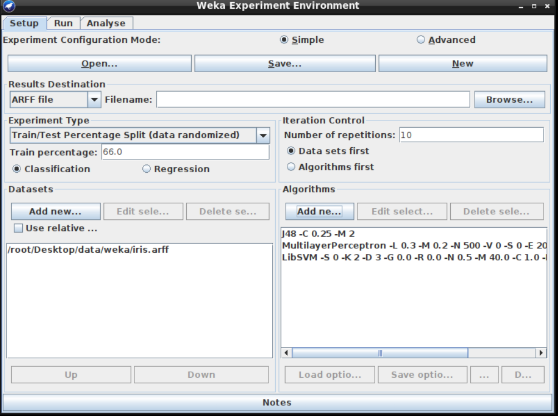
Description automatically generated with medium confidence

Train a support vector machine by using LibSVM program

**Question 5: What is the accuracy of the SVM classifier on the iris dataset? Capture the portion of classifier output that contains the accuracy and circle it. Then put it in your Lab report.**

The accuracy on the test data is 100%.

**6)**

****

Designing an experiment using requested dataset and algorithms

Graphical user interface, text, application, email

Description automatically generated

Analyse the result of experiment to rank the algorithms

Graphical user interface, text, application, email

Description automatically generated

Analyse the result of experiment to show the accuracy of algorithms

**Question 6: Which classifier is more accurate in classifying data in the iris dataset? What is the accuracy of the best classifier? Capture the portion of Test output that contains the accuracy of all three classifiers. Circle the name of the best classifier together with its accuracy. Then put it in your Lab report.**

According to the graph, since LibSVM has the highest accuracy than other two classifiers, which is 97.25%, it is the best classifier for the iris dataset.

Graphical user interface, text, application, email

Description automatically generated

**Part II: Clustering**

**Procedure:**

**1)**

Graphical user interface, application

Description automatically generated

Designing a cluster analysation with given parameter using EM

Graphical user interface, text, application

Description automatically generated

Evaluting the cluster data by assigning classes to it

Graphical user interface, scatter chart

Description automatically generated

Visualizing the cluster assignment and incorrectly clustered instances appear

**Question 7: Given that each class comprises 50 samples, which of the two classes contain the largest number of confusable data? Explain your answer. Note that if the cluster assignment is perfect, you should have seen the following results in the “Cluster output” window:**

**0 1 2 <-- assigned to cluster**

**0 50 0 | Iris-setosa**

**50 0 0 | Iris-versicolor**

**0 0 50 | Iris-virginica**

**Cluster 0 <-- Iris-versicolor**

**Cluster 1 <-- Iris-setosa**

**Cluster 2 <-- Iris-virginica**

Class Iris-setosa and Iris-virginica since according to the graph, there are 14 data records are confusable which they belong to Iris-setosa or Iris-virginica.

Graphical user interface, text, application

Description automatically generated

**2)**

Graphical user interface, text, application

Description automatically generated

Designing a cluster analysation with given parameter using K-Means

**Question 8: How many samples are wrongly clustered by K-means? How many samples are wrongly clustered by EM? Is K-means better or poorer than EM in clustering the iris data? Capture the portion of Clusterer output that contains the number of wrongly clustered samples by K-means and by EM. Circle the number and put it in your Lab report.**

There are 17 incorrectly clustered instances using K-means while there are only 14 incorrectly clustered instances using EM. Therefore, for this iris data, K-mean perform poorer than EM in clustering data.

Graphical user interface, text, application

Description automatically generatedGraphical user interface, text, application

Description automatically generated

Result generated by using (i) EM (ii) K-Means

**Part III: Regression**

**Procedure:**

**1)**

Graphical user interface, application

Description automatically generated

Open the WEKA explorer and select the cpu.arff file

Graphical user interface, text, application

Description automatically generated

Design a linear regression using given parameter

**Question 9: Which of the attributes does not affect the performance of the CPU? Explain your answer.**

Attribute CHMIN doesn’t affect the performance of the CPU since it doesn’t appear in the linear regression model. This indicates the rating of performance doesn’t depend on the CHMIN.