11/11/2020 **Mohd. Aftab Alam**
02013302717

EXPERIMENT 5

Aim: Implementation of Clustering technique on ARFF files using WEKA.

Theory:

This experiment illustrates the use of simple k-mean clustering with Weka explorer. The sample data set used for this example is based on the iris data available in ARFF format. This document assumes that appropriate preprocessing has been performed. This iris dataset includes 150 instances.

Steps involved in this Experiment

Step 1: Run the Weka explorer and load the data file iris.arff in preprocessing interface.

Step 2: In order to perform clustering, select the 'cluster' tab in the explorer and click on the choose button. This step results in a dropdown list of available clustering algorithms.

Step 3: In this case we select 'Simple K-Means'.

Step 4: Next click in text button to the right of the choose button to get popup window shown in the screenshots. In this window we enter six on the number of clusters and we leave the value of the seed on as it is. The seed value is used in generating a random number which is used for making the internal assignments of instances of clusters.

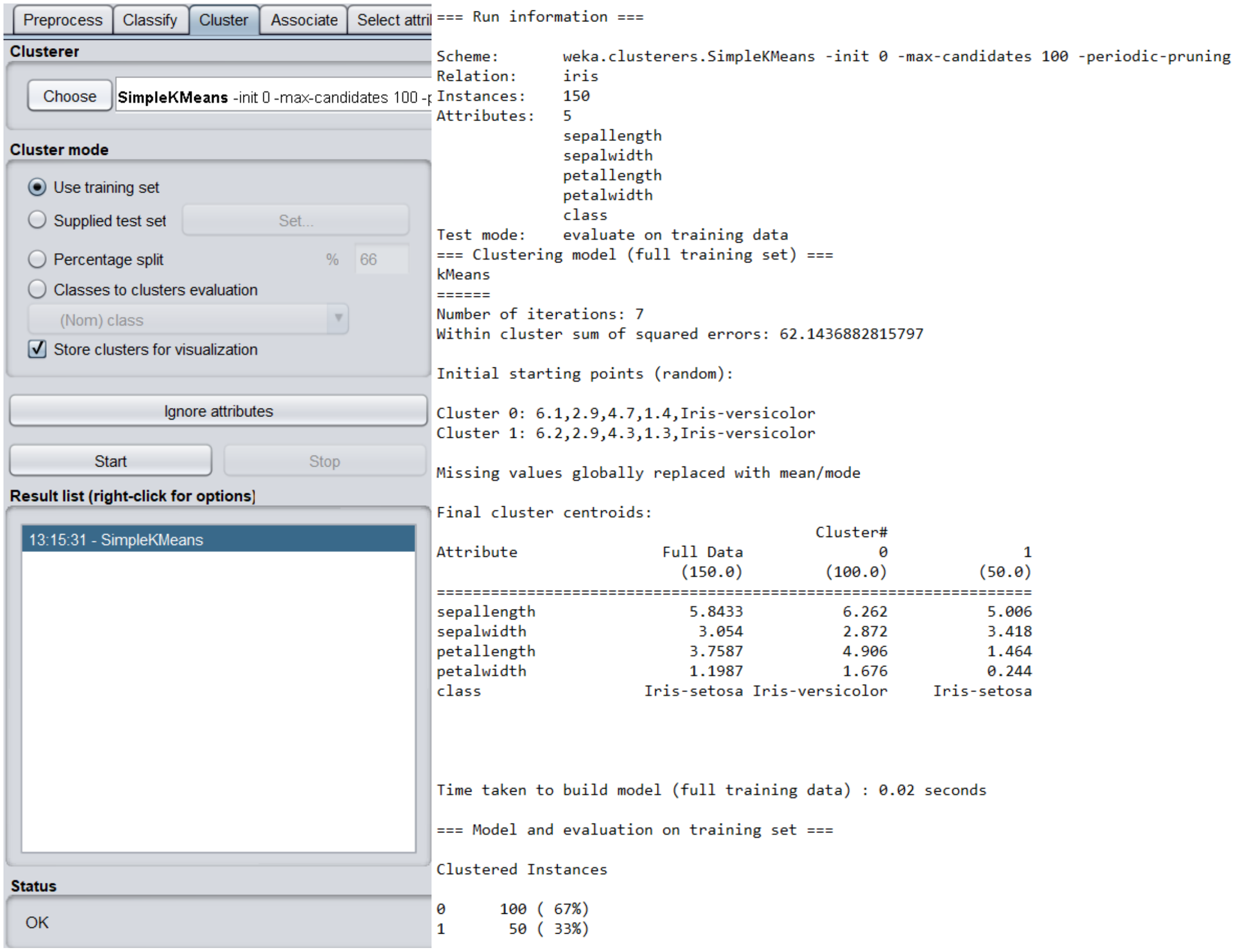
Step 5: Once of the option have been specified. We run the clustering algorithm there we must make sure that they are in the 'cluster mode' panel. The use of training set option is selected and then we click 'start' button. This process and resulting window are shown in the following screenshots.

Step 6: The result window shows the centroid of each cluster as well as statistics on the number and the percent of instances assigned to different clusters. Here clusters centroid are means vectors for each cluster. This clusters can be used to characterized the cluster. For e.g., the centroid of cluster1 shows the class iris. Versicolor mean value of the sepal length is 5.4706, sepal width 2.4765, petal width 1.1294, petal length 3.7941.

Step 7: Another way of understanding characteristic of each cluster through visualization, we can do this, try right clicking the result set on the result. List panel and selecting the visualize cluster assignments.

Step 8: From the above visualization, we can understand the distribution of sepal length and petal length in each cluster. For instance, for each cluster is dominated by petal length. In this case by changing the color dimension to other attributes we can see their distribution with in each of the cluster. We can assure that resulting dataset which included each instance along with its assign cluster. To do so we click the save button in the visualization window and save the result iris k-mean. The top portion of this file is shown in the following figure.

The following screenshot shows the clustering rules that were generated when Simple-K-Means algorithm is applied on the given dataset.



SimpleKMeans -init 0 -max-candidates 100 -p

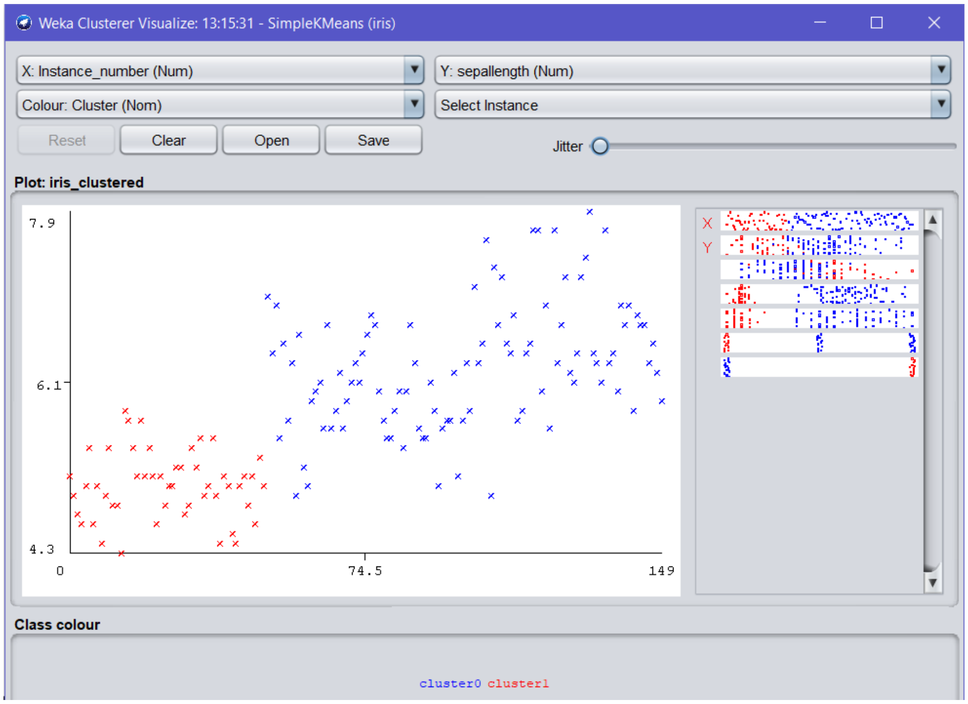
● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Nom) class
☑ Store clusters for visualization

13:15:31 - SimpleKMeans

OK

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:    evaluate on training data
=== Clustering model (full training set) ===
kMeans
======
Number of iterations: 7
Within cluster sum of squared errors: 62.1436882815797

Initial starting points (random):

Cluster 0: 6.1,2.9,4.7,1.4,Iris-versicolor
Cluster 1: 6.2,2.9,4.3,1.3,Iris-versicolor

Missing values globally replaced with mean/mode

Final cluster centroids:
                                        Cluster#
Attribute                 Full Data            0                1
                            (150.0)      (100.0)           (50.0)
==================================================================
sepallength                  5.8433        6.262            5.006
sepalwidth                    3.054        2.872            3.418
petallength                  3.7587        4.906            1.464
petalwidth                   1.1987        1.676            0.244
class                   Iris-setosa Iris-versicolor   Iris-setosa




Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      100 ( 67%)
1       50 ( 33%)
```

Visualization:



X: Instance_number (Num)

Y: sepallength (Num)

Colour: Cluster (Nom)

Plot: iris_clustered