COC131 Data Mining - Clustering

Martin D. Sykora m.d.sykora@lboro.ac.uk

Tutorial 05, Friday 20th March 2009

1. Fire up Weka (Waikako Environment for Knowledge Analysis) software, launch the explorer window and select the "Preprocess" tab. Open the iris data-set ("iris.arff", this should be in the ./data/ directory of the Weka install).

2. In week 2 (Tutorial 2) we looked at data preprocessing and visualisation, before we launch into Clustering its worth recapping the way in which data can be visualised using Weka. Figure 1 shows plots using instances of the iris data-set with possible class boundaries. (or *clusters*) marked on (you can see these plots in Weka under the "Visualisation" tab). Figure 1(a) shows a plot of attributes "petal width" vs "pepal length" and Figure 1(b) shows a plot of attributes "setal width" vs "sepal length". Each plot has new points drawn on marked 1 - 4. For each plot (a) and (b) and each point (1 - 4) which class does the new point belong to and why? Which plot is most reliable for class discrimination and why? In 1(b) the overlap D between class boundaries B and C is very large, do you think the class boundaries B and C are meaningful and why?

3. In Figure 1 determining the class boundaries manually was hard enough with labeled data (i.e. knowing the classes) however imagine a case where the classes are not known *a priori*. Look at Figure 2 this shows the same plots as Figure 1 but without labeling. Without the labeled data would you mark the same class boundaries as above? Which class boundaries would be particularly hard to place? What else do you notice about these plots?

4. Select the "Cluster" tab to get into the clustering perspective of Weka. Under "Clusterer" select and run each of the following clustering algorithms.
   - *cobweb, EM, farthest first, simple K-means*
   - under "Cluster mode" leave "Use training set" highlighted

5. Under the "Results list" you should now have five entries, right-clicking on an entry will give you the option to "Visualise cluster assignments". Inspect each of the entries for the plot (varying the x- and y- axes) and compare them to the class labeled plots (under the "Visualise" tab). How do the clusterers compare with the labeled plots? *The perfect clusterer should place all the instances of a particular class in the same cluster.* Try to change the numClusters parameter where possible - what happens to the cluster assignments?

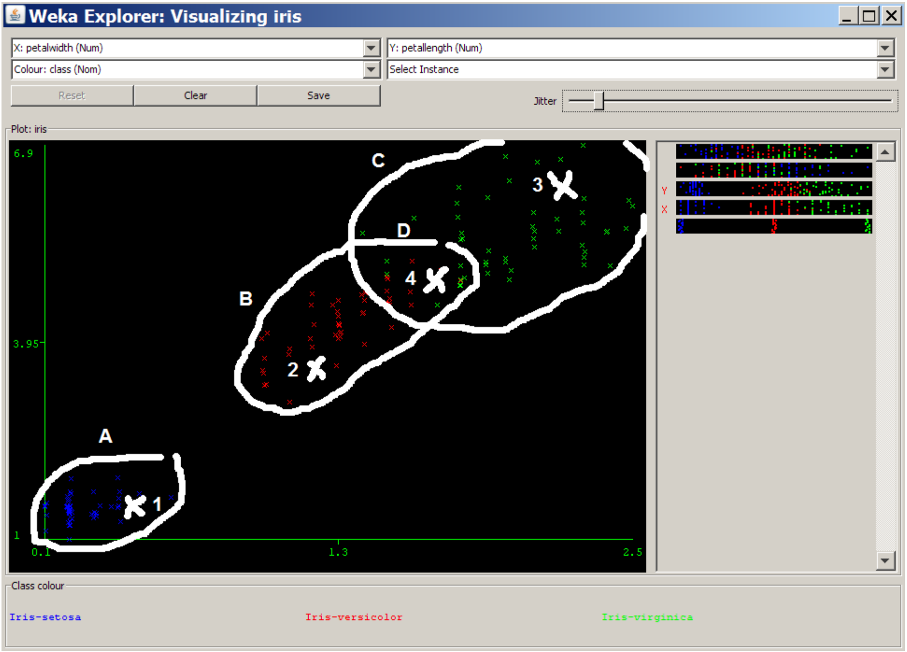
6. Now run the EM, farthest first, simple K-means clusterers again, but this time selecting under "Cluster mode", "Classes to clusters evaluation". In this mode Weka first ignores the class attribute and generates the clustering. Then during the test phase it assigns classes to the clusters, based on the majority value of the class attribute within each cluster. Then it computes the classification error, based on this assignment and also shows the corresponding confusion matrix.
   - For each clusterer inspect the confusion matrix and classification error. Which clusterer gives the best result?

7. Select the "Preprocessing" tab to get into the dataset loading perspective of Weka. Open the flag data-set ("flagdata.arff", this is available on the tutorial webpage http://www-staff.lboro.ac.uk/˜comds2). Go back to the clustering perspective of Weka by selecting the "Cluster" tab. Under "Clusterer" select and run the cobweb clustering algorithm, this is the only hierarchical clusterer within Weka.
   - make sure the two parameters acuity and cutoff are set to -A 1.0 -C 0.4 respectively. (They can be specified through the pop-up window that appears by clicking on area left to the Choose button.)
   - under "Cluster mode" leave "Use training set" highlighted

8. In the "Clusterer output" window we will notice a "tree" (dendrogram) generated by the clusterer. Do you understand what the tree is representing in terms of clusters? How many clusters are there in the first level of the tree? The tree can be ingterpreted as follows:
   - node N or leaf N represents a subcluster, whose parent cluster is N
   - The clustering tree structure is shown as a horizontal tree, where subclusters are aligned at the same column. For example, cluster 1 (referred to in node 1) has three subclusters 2 (leaf 2), 3 (leaf 3) and 4 (leaf 4).
   - The root cluster is 0. Each line with node 0 defines a subcluster of the root.
   - The number in square brackets after node N represents the number

of instances in the parent cluster N.
- Clusters with [1] at the end of the line are instances.
- Finally to view the clustering tree in graphical form right click on the last line in the result list window and then select Visualize tree.

Figure 1: Labeled data visualisation



Weka Explorer: Visualizing iris

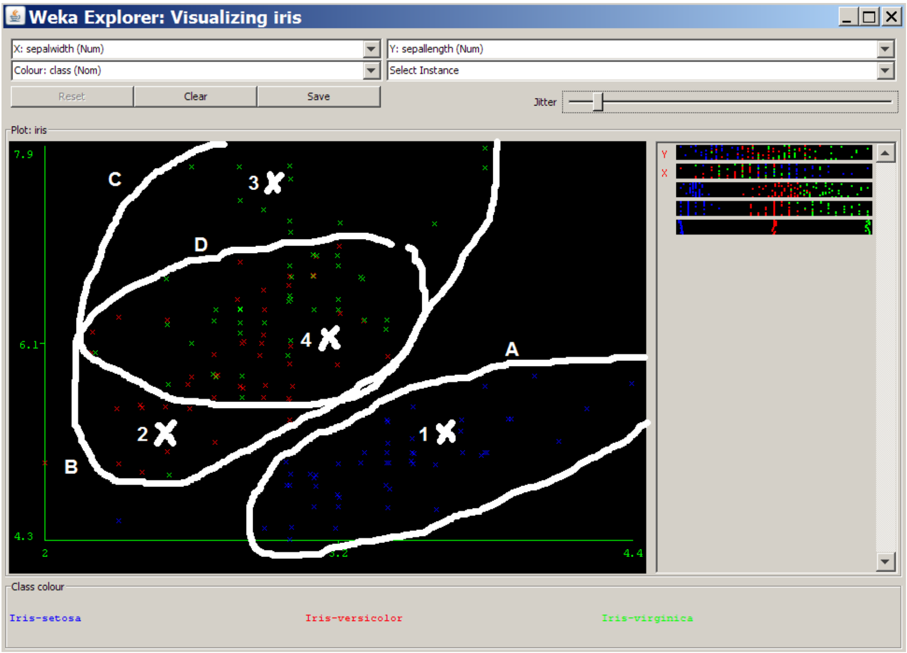
X: petalwidth (Num)

Y: petallength (Num)

Colour: class (Nom)

Select Instance

a



Weka Explorer: Visualizing iris

X: sepalwidth (Num)

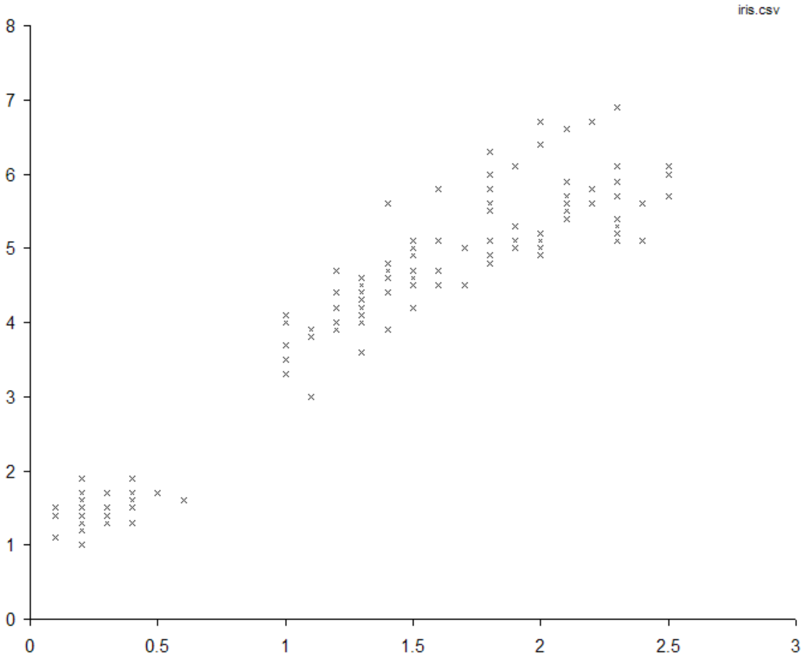
Y: sepallength (Num)

Colour: class (Nom)

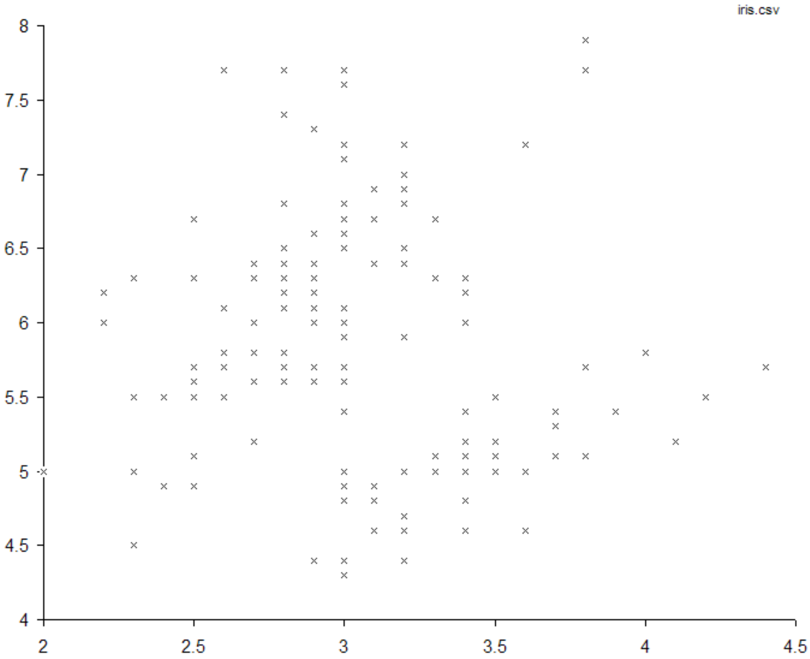
Select Instance

b

Figure 2: Unlabeled data visualisation



a



b