Lab1- Clustering

Given a dataset containing some information about 27 kinds of food, we are supposed to apply a clustering on that dataset implementing two different algorithms: K-means and density-based clustering.

The first step is to import the dataset in Weka. This can be done by the pressing “open file” in “preprocess” tab. This dataset contains 6 attributes or features including: Name, Energy, Protein, Fat, Calcium, Iron. “Name” contains the names of the 27 kinds of the food each of which has 5 values for the other features. Thus, in clustering we ignore the name as we interested in the measurements of the contents of these foods.

# K-means clustering

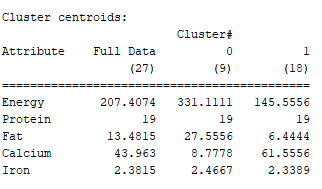
k-means is a centroid-based partitioning method which partitions the points into k partitions such that the highest possible within-cluster similarity and between-clusters dissimilarities are obtained. This method defines the centroid of a cluster as the mean value of the points within the cluster. Implementing this method require providing the algorithm two values: k or the number of clusters and the initial random selection of cluster centers or the mean values of those k partitions.

Pressing “choose” button in the “Cluster” tab enables us to select a method for clustering. We choose “SimpleKMeans”. By clicking on the name of the method next to “choose” button an editor is opened to set the parameters for the method. Here we can set the number of clusters(k) and the seed which is the random number initializing. For the similarity measurement we select the EuclideanDistance. Maximum number of iterations can also be set in this editor.

## Result for k = 2, seed =10

After just 2 iterations the algorithm stopped. In this case a value of 5.069 for within cluster sum of squared errors was obtained. The smaller the error, the highest the within-cluster similarity.

The table below summarizes the values of means for each cluster according to each feature. The centroid of the full data is 27 and of the clusters are 9 and 18. It seems that in cluster # 0 the values of Calcium are much closer to the mean value, and in cluster #1 Protein is much similar to the mean of the cluster. In both clusters Energy has the highest mean value. The lowest mean value belongs to Iron in both clusters.



In the visualize tab we can have the scatter plots of the attributes and investigate the clusters visually. Figure 1 shows a scatter plot in which the data points or food are plotted in a two-dimensional space. X-axis is the values of Energy and the Y-axis is of Protein. The two clusters (0 and 1) are clearly discernible. One possible interpretation may be as follow:

As the plot illustrates the foods may be divided into two groups: Low energy and high energy. Those with low energy have more variations of Protein content compared to the high energy foods.

Figure 2 illustrates such a scatter plot for Fat vs. Energy. A linear relationship with high correlation can be seen in this plot. As we expected the higher the fat a food contains, the higher the energy it will provide. In this plot the clusters are obvious as well.

For every other features we can do the same procedure to visualize the clusters in their space.

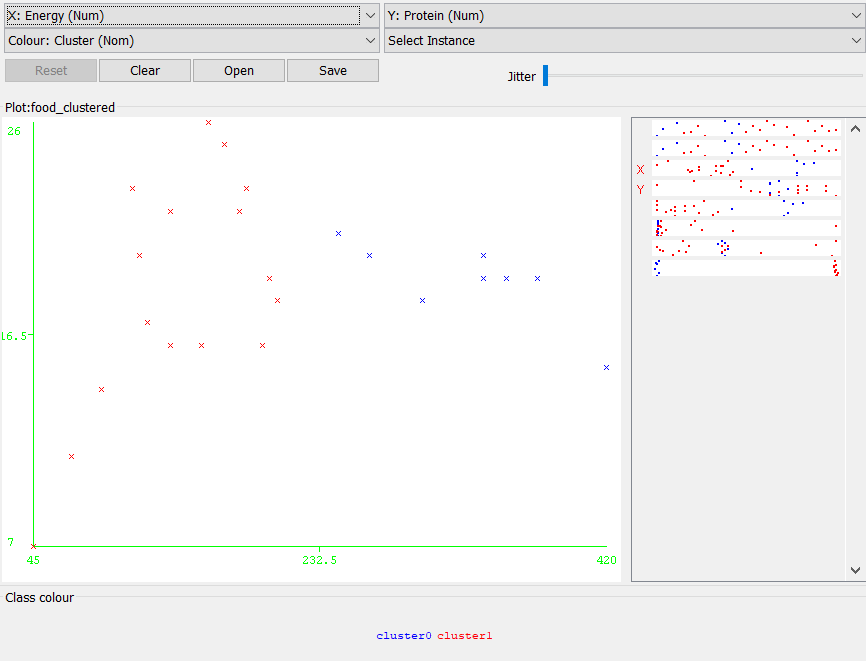


Figure 1. Scatter plot of Protein vs. Energy

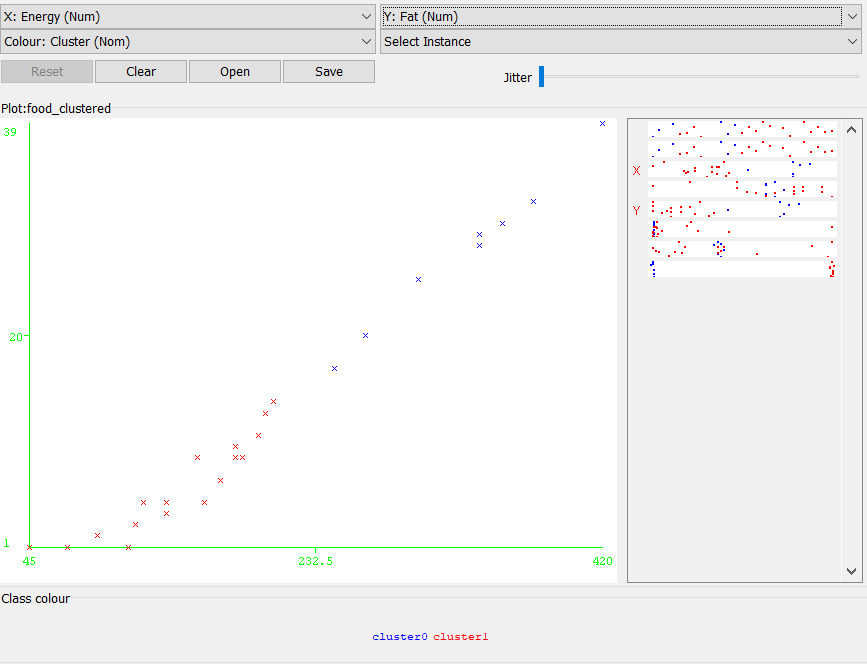


Figure 2. Scatter plot of Fat vs. Energy

## Result for k = 5, seed =10

In this case the algorithm did the clustering after 4 iterations and the within cluster sum of squared errors was resulted as 2.7. It seems the performance increased.

The following table summarizes the centroid values for the full data and for each cluster.

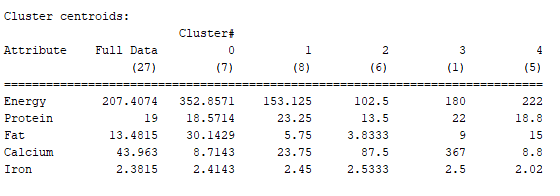


Figure 3 shows the same scatter plot as figure 1. In this case the points have been divided into partitions according to their Protein contents as well as their energy contents.

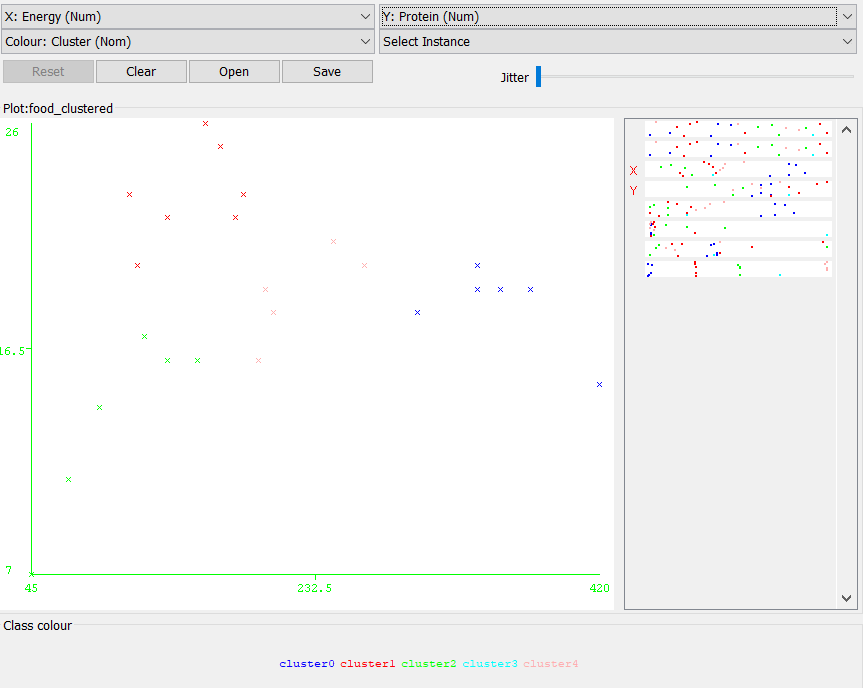


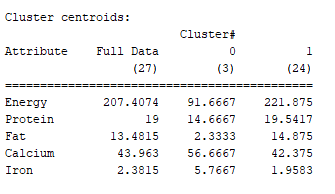
Figure 3. Scatter plot of Protein vs. Energy

It seems that increasing the number of partitions enabled us to analyse the foods in more details.

## Result for k = 2, seed =20

In this case the within cluster errors is 6.45 which is higher than the error obtained with seed of 10.

The table of centroids summary is as follow:



In comparison to the case with k=2 and seed =10, we can see that the values of the centroids for the clusters have been changed. Seed controls the randomly selection of the mean values for the clusters in order to initializing the algorithm. Hence different seeds result in different initial values for the means and so the final results may vary.

Figure 4 is the scatter plot of Protein vs Energy. As it can be seen the result is not good. The clusters are not produced properly. This is because of bad selection of the initial values for the mean.

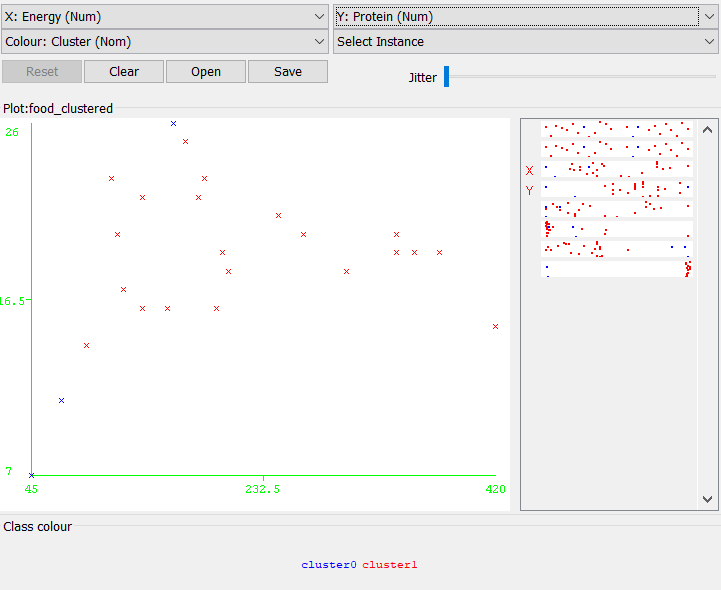


Figure 4. Protein vs. Energy