Clustering

Lab 2: Illustration of clustering techniques with real datasets

- Partitional clustering
- Quality indexes
- Hierarchical clustering

Please send comment to (version 1.0 – January 6, 2017):

Romain Billot *romain.billot@telecom-bretagne.eu*
Yannis Haralambous *yannis.haralambous@telecom-bretagne.eu*
Philippe Lenca *philippe.lenca@telecom-bretagne.eu*
Sorin Moga *sorin.moga@telecom-bretagne.eu*

1 Data presentation

–The Ruspini dataset– The first data we are going to use is the Ruspini data set, a two-dimensional dataset traditionnally used to illustrate clustering partional techniques:

Variable	Meaning
`x`	first coordinate
`y`	second coordinate

Table 1: Variables of the ruspini dataset

–The food dataset– The second data contains information about food. Nutrient levels were measured in a 3 ounce portion of various foods.

Variable	Meaning
`Name`	name of the item
`Energy`	number of calories
`Protein`	amount of protein in grams
`Fat`	amount of fat in grams
`Calcium`	amount of calcium in milligrams
`Iron`	amount of iron in milligrams

Table 2: Variables of the `food.csv` dataset

Question 1

Load and describe the 2 datasets into R. Thanks to boxplots, try to highlight some outliers, if any.

2 Partitional clustering and quality of a clustering

For the illustration of the algorithms seen during the lecture class, we will use the Ruspini dataset.

Question 2

Plot the data in a 2D space. What can you say about the structure of the data? How many cluster do you expect to be found by the clustering algorithms?

Question 3

Run a k-means algorithm with a given number of clusters. Plot the ruspini data set and assign to each point a specific color according to its group. Repeat the same operation several times while vizualizing the results. What do you remark ?

Question 4

Try to tune some parameters found in help(kmeans) in order to fix the issue seen before.

Question 5

Determining the optimal number of clusters is a critical issue in clustering. In many applications, in the real life, this optimal number is not obvious and we use quality indexes that will evaluate the quality of a clustering. For example, some quality indexes are based on the idea that a good clustering is composed of compact and well-separated clusters. The idea is then to launch several runs of a clustering algorithm with various initial numbers of clusters. The optimal number of cluster is the one for which the quality index is maximized. In this question we illustrate this procedure with the silhouette index that you can find in the FPC package. You are asked to write a code that will compute the silhouette for a k-means clustering of the Ruspini dataset with a initial number of clusters ranging from 2 to 10. Plot the evolution of the silhouette index against the initial number of clusters. What is your conclusion ?

Question 6

The PAM algorithm stands for Partitioning Around Medoïds. It is a more robust version than the classic kmeans algorithm. Verify it by applying a PAM procedure to our Ruspini data set (cluster package). What can you conclude about the robustness or stability of the algorithm? Explore the features of the cluster package, for example by plotting the object resulting from the pam clustering

3 Hierarchical clustering

In this section, the food dataset is used.

Question 7

Have a first look at this dataset and try to make a partitional algorithm procedure like before.

Question 8

Perform a hierarchical clustering by first standardizing the colums of interest, then computing a dissimilarity between the objects, and finally running an agglomerative algorithm to yield a dendrogram. Interpret the results.

Question 9

Try to hightlight the impact of the chosen distance on the results. Next, try to show the impact of the agglomerative criteria on the results

Question 10

Intrepret a clsutering of 4 groups by cutting the dendrogram at a given number of classes. You can use the cutree function to cut the dendrogram and interpret the groups with descriptive statistics. As an agglomerative criterion, you may prefer the ward method in order to get well-balanced groups.

Home work

At home, before sleeping, you can have fun and plot nice dendrograms. You can visit this webpage and try to do the same with our food dataset.
http://www.sthda.com/english/wiki/beautiful-dendrogram-visualizations-in-r-5-must-known-methods-unsupervised-machine-learning