University of Ferrara

DIPARTIMENTO
DI ECONOMIA
E MANAGEMENT

Statistics for Economics and Business

Stefano Bonnini & Valentina Mini

Cluster Analysis:
non hierarchical methods

Lecture 11 – 27nd of March 2019

NHCA - introduction

“Data mining methods are needed to obtain
useful information for users in electronic environment.

One of these methods, clustering methods, aims to group
data according to common properties.

Clustering methods are divided in two categories.
These are hierarchical clustering and **non-hierarchical clustering methods.**

Non-hierarchical clustering methods are divided in four sub-classes:
a) partitioning,
b) density-based,
c) gridbased
d) and other approaches”

(Gulagiz F.K and Sahin S. (2017) *Comparison of Hierarchical and Non Hierarchical Clustering Algorithms*, International Journal of Computer Engineering and Information Technology January 2017, 6-14 (available online))

NHCA - introduction

The algorithms of the hierarchical class gather the most similar two objects in a cluster. This has very high process cost because all objects are compared before every clustering step.

Clustering algorithms in non-hierarchical category cluster the data directly.
Such algorithms generally change centers until all points are related to centers.

1) **K-Means** algorithms are low cost in terms of calculation time (compared with HCA): they are based on the centroids calculation.

2) **Density based** clustering approaches consider intensive data spaces as cluster, so have no problem in finding clusters with random shapes. Among the best known density based clustering algorithms: DBSCAN (Density-based spatial clustering of applications with noise) and OPTICS (Ordering points to identify the clustering structure) algorithms.

3) **Grid based** clustering approach takes into consideration the cells rather than data points. Because of this feature, grid based clustering algorithms seems to be generally more effective as all computational clustering algorithms

NHCA - introduction

In contrast to the hierarchical method, these partitioning techniques **permit objects to change group membership through the cluster formation process**.

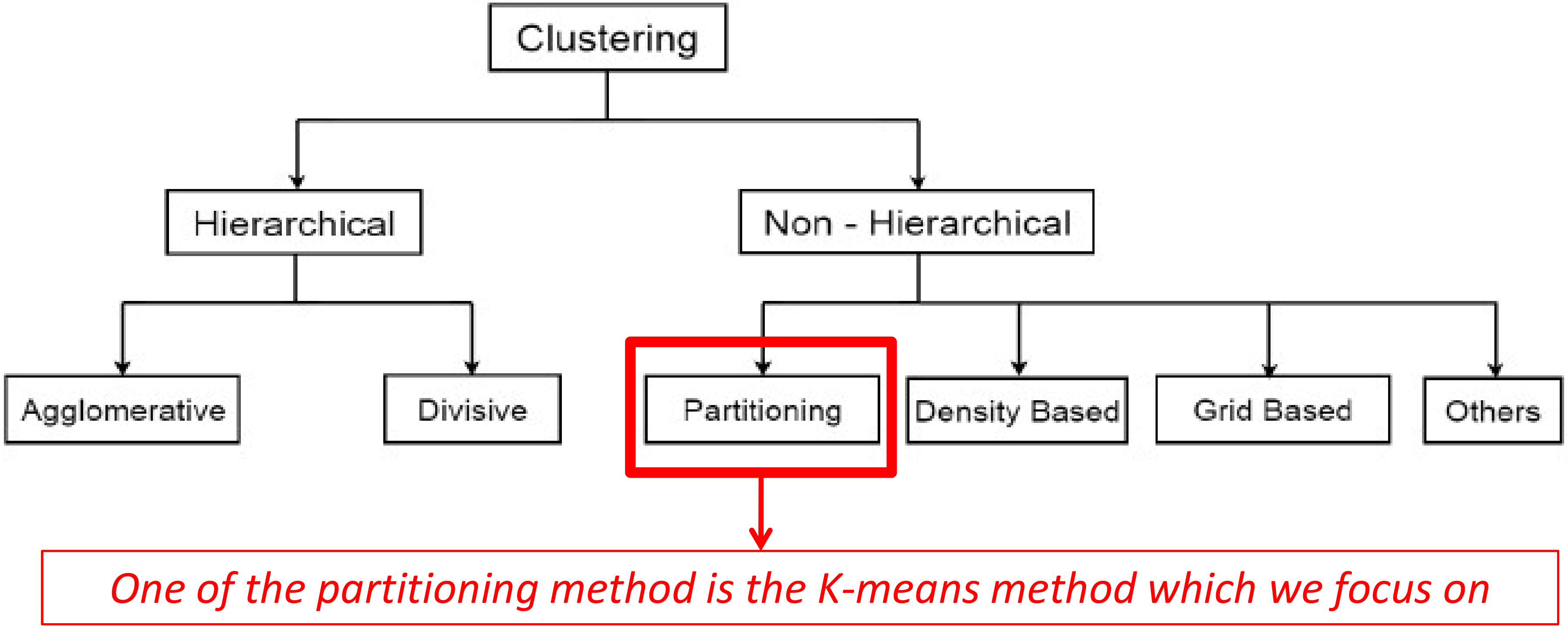
The partitioning method usually begins with an **initial solution**, after which reallocation occurs according to some optimality criterion.

Partitioning method constructs **g** clusters from the data as follows:

-Each clusters consists of at least one **object n** and each object must be belong to one clusters. This condition implies tha **g<=n**

- The different clusters **cannot have same object**, and the construct

NHCA - introduction



One of the partitioning method is the K-means method which we focus on

It is the most used method, because K-Means algorithm is less complex than other methods and the implementation of the algorithm is easier.

NHCA: K-means

K-Means algorithm consists of four basic steps:

- Determination of centers.

- Assigning points to clusters which are outside of the centers according to distance between centers and points.

- Calculation of new centers.

-Repeating these steps until obtaining decided clusters.

The biggest problem of K-Means algorithm is determination of starting points.

NHCA: K-means

With the **non hierarchical methods** we get just one partition of the ***n* units into *g* clusters**, for a **pre-determined number *g* of clusters**

The **rule** for the allocation of the units into the clusters considers an **objective function**, usually based on the **breakdown of the total deviance** (*TD*)

NHCA example

- *Example* - Wine survey on Passito:
  - Starting partition:

    customers are classified into ***g* groups**

  - Intermediate partitions:

  the customers are **reallocated in the groups** and, for each reallocation, the corresponding value of the **objective function is computed**

  Each customer is **assigned to a new group** when this assignment provides the greatest **improvement of internal cohesion**

  - The reallocations are **repeated** until a **given stopping rule** is satisfied

NHCA: K-means

- Weaknesses of the procedure:

(1) the choice of the number g of groups is arbitrary;

(2) the starting partition affects the final result

NHCA: K-means

This method is developed by Queen (1967).
He suggests the name K-means for describing his algorithm that **assigns each item to the cluster having the nearest centroid (mean).**

This process consists of following steps:
1.Partition the items into g initial clusters
2.Proceed through the **list of items**,
3.assigning an item to the cluster whose centroid (mean) is nearest.
Recalculate the centroid for the cluster receiving in the new item and
for the cluster losing the item.
4.Repeat step 2 and 3 until no more assignments take place

It is better to determine g initial centroids (seed points) first, before proceeding to step 2.

This method **tries to minimize the sum of the within-cluster variances (WD)**.

NHCA: K-means' steps

1. Chose ***g* starting seeds or poles** as centroids of the starting partition **and assign each unit** to the cluster with the **nearest centroid**

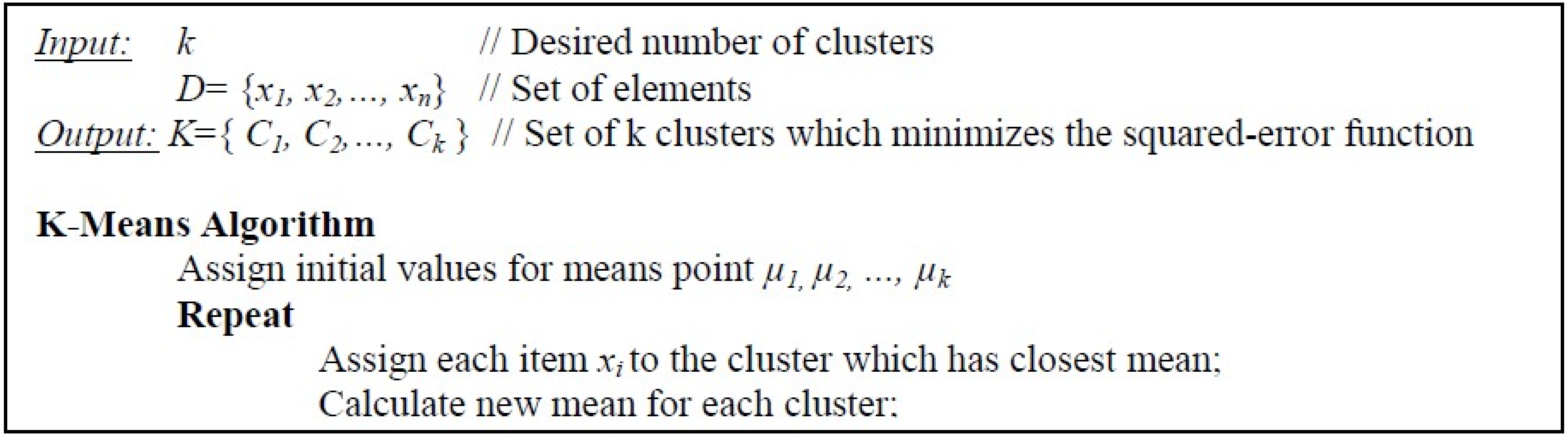
2. Compute the **centroids of the *g* new clusters** created at step (1)

3. Assign each unit to the new cluster **with the nearest centroid**

4. **Repeat step (2) and step (3) until one of the following convergence rules is satisfied:**
    I. R^2 variation is less than a given treshold
    II. The changes of the centroid positions are less than a given treshold
    III. The number of iterations reaches a certain predetermined value
    IV. …

Remark: with Euclidean distance we always have convergence of the algorithm

NHCA: K-means algorithm



```
Input:   $k$                              // Desired number of clusters
         $D= \{x_1, x_2, \dots, x_n\}$    // Set of elements
Output: $K=\{ C_1, C_2, \dots, C_k \}$   // Set of k clusters which minimizes the squared-error function

K-Means Algorithm
    Assign initial values for means point $\mu_1, \mu_2, \dots, \mu_k$
    Repeat
        Assign each item $x_i$ to the cluster which has closest mean;
        Calculate new mean for each cluster;
```

Lab using R

- Read the eating database
- Perform a non hierarchical cluster analysis to detect 3 main clusters of countries based on their vitamins assumption (please take into account the veg-pulses and fruits consumption)
- Comment your results

R “at home” exercises

Problem 1 - Passito

- Perform a hierarchical CA on the 17 response variables of the questionnaire which represent habits, behaviors and preferences of wine drinkers (from variable LIKE_WINE to variable PRICE) to detect homogeneous market segments of wine drinkers

- Perform a k-means CA on the 17 response variables of the questionnaire to detect 4 homogeneous market segments of wine drinkers