Università degli studi di Ferrara
Dipartimento di Matematica
A.A. 2019/2020 – I semestre

STATISTICA MULTIVARIATA

SSD MAT/06

LEZIONE 16– Metodi non gerarchici per l'analisi di clusters

Docente: Valentina MINI

valentina.mini@unife.it

RICEVIMENTO: LUNEDI POMERIGGIO, appuntamento previa mail

NHCA - introduzione

I metodi di raggruppamento sono suddivisi in due aree: cluter gerarchici e clutsre non gerarchici.

I metodi non gerarchici sono suddivisi in 4 principali categorie:
a) partitioning,
b) density-based,
c) gridbased
d) and “other approaches”

Lettura consigliata:
(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))

The state of the art

CA: aim = identify the lower number of clusters such that

The units belonging the same cluster are more similar than …→	The units belonging different clusters
High within-cluster similarity	Low between-cluster similarity
Low within-cluster variance	High between-cluster variance

To identify clusters we should define

Distance or similarity

Distance:

-Euclidean

-Manhattan

-Minkosky

-Chebichev

Similarity:

1. case of dichotomous var.
2. case of categorical var.

:

*Ind. of co-presences (Russel&Rao; Jaccart)

*Ind. Co-presences and co-absences (Sokal & Michener)

Grouping's rule

Hierarchical methods

Non Hier. methods

Divisive:

-Edwards & Cavalli Sforza (trace of the deviance matrix)

-Friedman &Rubin (min. the deviance matrix determinant)

Agglomerative:

-Single linkage
-Complete linkage
-Average linkage

•Centroide method
• Ward method

NHCA - introduzione

L'obiettivo dell'algoritmo è quello di inserire due oggetti simili nel medesimo gruppo
Questo processo del cluster gerarchico ha un alto costo in quanto tutti gli oggetti sono confrontati prima di ogni step.

INVECE
L'algoritmo nel cluster nonn gerarchico raggruppa i casi direttamente, modificando il punto centrale fino a quando la struttura dei gruppi non cambia.

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 - introduzione

Queste tecniche permettono ai casi di modificare la loro appartenenza al gruppo durante il processo di allocazione.

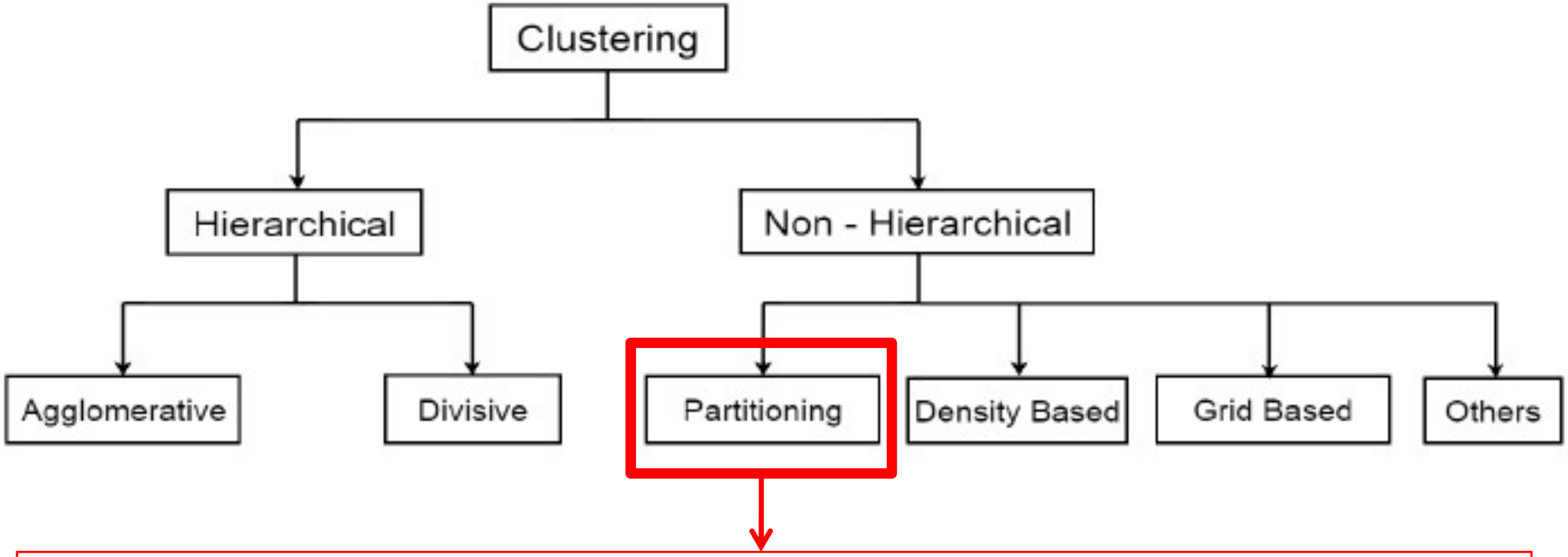
Il metodo di partizione (partitioning) generalmente inzia con una SOLUZIONE INIZIALE, dopo la quale avvengono molte riallocazioni seguendo il criterio di ottimo.

Il metodo non gerarchico costruisce **g** clusters dai dati seguendo il processo:

-Ogni cluster consiste di almeno un oggetto **n** e ogni oggetto deve appartenere ad un solo cluster. Questa condizione implica che **g<=n**

- Cluster diversi non possono avere stessi oggetti

NHCA - introduzione



Uno dei principali metodi non gerarchici è K-Means

K means è uno dei metodi più utilizzati in quanto il processo di allocazione (algoritmo) è meno complesso rispetto ad altri e l'implementazione dello stesso è più semplice.

NHCA: K-means

K-Means algorithm si basa su **4** steps:

1. Determinazione dei centroidi
2. Assegnazione dei casi ai clusters in base alla distanza tra ogni caso e il centroide.
3. Definizione dei nuovi centroidi.
4. Ripetizione di queste fasi fino a quando non si ottiene la soluzione migliore e stabile.

Il prinicpale problema del K-means è il fatto di dover definire a priori il numero di cluster.

NHCA: K-means

Con I metodi non gerarchici si ha una partizione delle n unità in g cluster **sulla base di un numero predeterminato di g.**

La regola per l'allocazione delle unità statistiche ai cluster si basa su una funzione oggettiva, generalmente identificata dalla breakdown della deviaza totale (TD)

NHCA esempio

- *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

- Debolezze della procedura:

(1) La scelta del numero di gruppi g è arbitrario;

(2) La **partizione iniziale** determina il risultato finale

NHCA: K-means

Questo metodo è stato sviluppato da Queen (1967).
Ha suggerito il nome K-means per indicare che l'algoritmo assegna ogni caso al cluster che ha il più vicino centroide **(mean).**

Questo processo consiste di 4 step:
1. raggruppamento dei casi in g gruppi
2.Si procede su tutta la lista dei casi,
3.Si riassegna ogni caso al gruppo il cui centroide è più vicino.
Si ricalcolano I centroidi e le distanze dei casi da essi.
4.Ripetere le fasi 2 e 3 fino a quando nessuna nuova assegnazione deve essere effettuata

Questo metodo cerca di minimizzare la somma delle varianze interne ai clusters (WD)

NHCA: algoritmo K-means

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

K-Means Algorithm

Assign initial values for means point $\mu_1, \mu_2, \ldots, \mu_k$
Repeat
Assign each item x_i to the cluster which has closest mean;
Calculate new mean for each cluster;

Lab con R

R exercises

- Dataset sulle abitudini alimentari
- NHCA per identificare 3 gruppi principali di paesi sulla base della loro assunzione di vitamine (please take into account the veg-pulses and fruits consumption)
- Commentate I risultati

R exercises

Problem 1 - Passito

- Effettuare una HCA sulle 17 variabili risposta del database Passito (dalla variabile LIKE_WINE a PRICE) per identificare segmenti omogenei di mercato relativi ai consumatori di vino

- Effettuare una NHCA (K-means) sulle 17 variabili risposta per identificare 4 segmenti omogenei del mercato relativi ai consumatori di vino