Analisi dei dati: Analisi dei gruppi

Domenico De Stefano

a.a. 2022/2023

Indice

Cosa intendiamo per "classificare"

- L'obiettivo generale di una procedura di classificazione è, come dice il nome, suddividere i dati in classi (gruppi).
- Più precisamente si vuole, in presenza di individui (unità statistiche) di cui si conoscono determinate caratteristiche (variabili), costruire dei criteri per assegnare ciascun individuo a un gruppo sulla base delle caratteristiche note.
- Distinguiamo due casi secondo la natura dei gruppi:
  - se il fatto gli individui siano suddividibili in gruppi è un dato dell'indagine si parla di classificazione supervisionata.
  - se l'esistenza di una 'sensata' partizione degli individui in gruppi non è scontata, ma anzi è uno degli scopi dell'indagine decidere se esistano dei gruppi si parla di classificazione non supervisionata o raggruppamento

Tipo di dati: caso non supervisionato

Disponiamo di osservazioni relative a p caratteristiche di n individui. Formalmente abbiamo cioè una matrice di dati

$$X = \begin{bmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix} = (\mathbf{x}_1, \dots, \mathbf{x}_p),$$

- x_{ij} è l'osservazione della j-ma variabile sull' i-mo soggetto;
- colonna $(x_{1j}, \dots, x_{ij}, \dots, x_{nj})$: osservazioni della variabile j;
- riga $(x_{i1}, \dots, x_{ij}, \dots, x_{ip})$ caratteristiche dell'individuo i.

Tipo di dati: caso non supervisionato

Si consideri un campione costituito da quattro individui di cui si osservi l'età, il sesso e il titolo di studio:

Individuo	Genere	Età	Titolo di Studio
I	M	27	laurea
II	F	31	laurea
III	F	21	diploma
IV	M	19	diploma

Si ha dunque $p = 3$ e $n = 4$; la matrice X è costituita dalle ultime tre colonne della tabella, si ha poi $\mathbf{x}_1 = (M, F, F, M)$; $\mathbf{x}_2 = (27, 31, 21, 19)$. Si noti come le variabili abbiano natura diversa, categorica non ordinata, categorica ordinata, numerica.

Tipo di dati: caso supervisionato

Disponiamo, oltre alle suddette variabili, anche della classe di appartenenza di ciascuno degli individui, determinata, appunto da un 'supervisore', da cui il nome. (Si suppone cioè che la classe di appartenenza di un individuo sia osservabile direttamente.)

$$(\mathbf{y}, X) = \begin{bmatrix} y_1 & x_{11} & \dots & x_{1j} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots & & \vdots \\ y_i & x_{i1} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ y_n & x_{n1} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix} = (\mathbf{y}, \mathbf{x}_1, \dots, \mathbf{x}_p).$$

- $y_i \in \mathcal{G} = (G_1, \dots, G_g)$ è la classe di appartenenza dell'individuo i-mo.

Tipo di dati: caso supervisionato

Riprendiamo i dati dell'esempio precedente e aggiungiamo la variabile risposta Y che rappresenta lo stato occupazionale, occupato/non occupato, sicché la matrice dei dati è completata come

occupato	M	27	laurea
occupato	F	31	laurea
non occupato	F	21	diploma
occupato	M	19	diploma

dunque $\mathbf{y}$ =(occupato, occupato, non occupato, occupato) mentre $\mathcal{G}$ =(occupato, non occupato).

Motivazioni: raggruppamento

- La classificazione non supervisionata ha lo scopo più ambizioso, si cerca infatti di determinare una suddivisione degli individui in gruppi sulla base delle caratteristiche osservate
  - senza sapere se una suddivisione in gruppi abbia senso e, di conseguenza,
  - senza sapere quale sia il numero dei gruppi.
- Si utilizza ad esempio da parte delle aziende per la segmentazione del mercato, per distinguere cioè tipologie di clienti diversi (al fine ad esempio di inviare pubblicità mirata).

Motivazioni: classificazione

- La classificazione supervisionata trova applicazione quando la 'supervisione' è un metodo inadeguato per classificare nuove unità: si cerca quindi un criterio di decisione basato sulle altre caratteristiche osservabili.
- L'inadeguatezza della 'supervisione' può riferirsi a circostanze diverse: nelle diagnosi di malattie quando la supervisione richiede un esame post mortem, nel riconoscimento ottico dei caratteri perché si vuole appunto tagliar fuori il supervisore, nel rischio di credito dove si vuol stabilire la rischiosità prima di prestare i soldi.
- In molti casi oltre allo scopo previsivo – di sostituzione del supervisore – vi può essere uno scopo descrittivo (i risultati della classificazione indicano rispetto a quali variabili si differenzino maggiormente i gruppi).

Metodi

- I metodi in uso per costruire un criterio di classificazione sono di natura diversa ed è difficile darne una descrizione comune, in termini vaghi si può dire che si cerca di riassumere le variabili esplicative in modo da esaltare quelle caratteristiche che differenziano maggiormente l'appartenenza a un gruppo piuttosto che a un altro.
- Per quanto riguarda la classificazione non supervisionata si considereranno cenni alle tecniche di *clustering* (raggruppamento)

Risultato

Il risultato di un'analisi di raggruppamento è

- una partizione dello spazio generato dalle p variabili esplicative (che, nel caso della classificazione non supervisionata, è lo spazio campionario).

Indice

Perché raggruppare

Suddividere collezioni di cose in gruppi è un modo naturale e, si può dire, imprescindibile, di ragionare per comprendere i fenomeni.
Si ragiona per gruppi perché è più facile dominare mentalmente pochi gruppi che tante unità.
Uno stesso insieme di enti consente diversi raggruppamenti, nessuno è 'giusto', semmai può essere utile (o inutile o anche dannoso).
Raggruppare utilmente: mettere insieme unità simili e separare unità dissimili, in altre parole creare gruppi omogenei al loro interno e disomogenei tra di loro (*internal cohesion and external isolation*).

Perché raggruppare (2)

Un raggruppamento non è necessariamente una suddivisione in gruppi *tout court*, si parla di raggruppamento gerarchico quando si considerano successive categorizzazioni via via più raffinate.
(Una partizione è più fine di un'altra se i suoi costituenti sono sottoinsiemi dei costituenti di quella.)
Nell'ambito scientifico questo tipo di categorizzazioni sono le più frequenti, ad. es. la classificazione dei viventi in biologia, in cui sulla base delle caratteristiche (morfologiche e altro) si formano gruppi via via più ristretti, in tal caso la gerarchizzazione ha anche un significato dal punto di vista evolutivo.

Perché raggruppare (3)

La categorizzazione può essere un importante strumento di ragionamento, al punto di suggerire ipotesi o campi di ricerca, la sua costruzione è però complessa e si rischia che essa sia fuorviante se eccessivamente semplificatrice.
L'analisi di raggruppamento è uno strumento che può aiutare nel processo di categorizzazione nel senso di suggerire una possibile suddivisione degli oggetti dell'analisi.
In particolare, l'analisi di raggruppamento (o analisi dei *cluster*) ha per scopo far emergere dall'insieme dei dati a disposizione gruppi di unità statistiche simili tra loro e dissimili da quelle degli altri gruppi dove la somiglianza è definita in qualche modo in base a una distanza.
Nell'analisi di raggruppamento la domanda cui si vuol rispondere è se esistono e quanti sono dei gruppi sensati (naturali) in cui suddividere le unità sulla base delle variabili osservate.

Esempi di applicazione (Kettenring, 2006)

- Biologia: raggruppamenti dei viventi (tassonomia) usando caratteristiche morfologiche o altro.
  - Parker et al. (2004) studiano i raggruppamenti ottenibili sulla base delle caratteristiche genetiche di cani di razza per confrontarli poi con i gruppi costituiti, appunto, dalle razze. L'interesse è verificare il grado di separazione genetica tra razze canine.
  - Kim et al. (2003) costruiscono raggruppamenti usando 25 caratteri quantitativi di piante acquatiche provenienti da 4 specie riconosciute da cui verificare la corrispondenza del raggruppamento statistico con quello a priori.
- Statistica sanitaria: valutazione della comorbidità. John et al. (2003) costruiscono gruppi di individui sulla base delle malattie diagnosticate a ciascuno, nei gruppi di individui che emergono sono tipici certi gruppi di malattie e sono quindi messe in luce associazioni tra le stesse.

Esempi di applicazione (Kettenring, 2006) (2)

- Psicologia: l'analisi dei gruppi è sovente utilizzata, ad esempio
  - Allik et al. (2003): tentano una classificazione di 36 culture (nazionalità) sulla base delle risposte di individui ad esse appartenenti a un questionario sulla personalità, ottenendo dei gruppi che, almeno in una certa misura, riflettono la geografia.
  - Stefurak et al. (2004) vogliono determinare i tratti di personalità prevalenti tra i delinquenti giovanili e individuare di conseguenza le categorie cliniche rilevanti.
- Fruizione dei media: van Rees et al. (2003) si pongono l'obbiettivo di individuare dei profili di utilizzo di diversi (19) media nella popolazione olandese usando il tempo speso da ciascuno di 28000 individui su ciascun medium, emerge una suddivisione in cui, ad esempio, sono associati quotidiani di qualità e video in un gruppo mentre in un altro gruppo si trovano giornali locali, radio e televisione.

Esempi di applicazione (Kettenring, 2006) (3)

- Archeologia: il settore archeologico è un ambito fertile per l'AG, ad esempio Hall (2004) considera la composizione di diversi esemplari di vasellame provenienti da sei siti, individua così dei gruppi che corrispondono a gruppi di siti e ciò suggerisce l'utilizzo comune a più siti di determinate fonti di materia prima e/o frequenti scambi tra i siti raggruppati insieme.

Raggruppamento gerarchico e non

Tecnicamente, si distinguono i metodi di partizione e quelli gerarchici.
Nei metodi di partizione, scelto un numero k di classi, si determina una partizione in k elementi.
Nei metodi gerarchici si individua una sequenza di partizioni nidificate (via via più fini) da quella in un unico elemento a quella con tenti elementi quante sono le osservazioni.
Da un metodo gerarchico posso sempre ottenere una partizione di k elementi per qualunque k, si noti però che la determinazione di questa dipende anche dalle suddivisioni precedenti (se siano le più fini o le meno fini dipende dall'algoritmo usato), un metodo di partizione non prende in considerazione suddivisioni in un numero di elementi diverso da k.

Indice

Raggruppamento gerarchico e non

Supponiamo di voler suddividere le unità in g gruppi ($g << n$, in generale). Supponiamo di aver determinato una funzione obbiettivo rispetto alla quale confrontare le partizioni alternative, ad es. sia

$$f(\cdot) = \frac{\text{Var. interna ai gruppi}}{\text{Var. tra i gruppi}}$$

sceglieremo in tal caso la partizione che minimizza la funzione obbiettivo. Della funzione obbiettivo si può dire, in generale, che essa risulterà funzione decrescente della coesione all'interno dei gruppi e della distanza tra gruppi diversi.

Raggruppamento gerarchico e non (2)

A questo punto un'opzione è creare tutte le possibili partizioni delle n unità in g gruppi e confrontarle rispetto a $f(\cdot)$, si può pero fare solo per valori molto bassi di n e g poiché il numero di partizioni di n unità in g gruppi diventa, al crescere di n e g, rapidamente molto elevato.
(In particolare, il numero è

$$N(n,g) = \frac{1}{g!}\sum_{k=1}^{g}\binom{g}{k}(-1)^{g-k}k^n$$

come riportato in Rencher (2002) p. 455. Per $n = 20$ e $g = 4$ si ottiene un numero di partizioni vicino a 45 miliardi).
Si opera dunque usando algoritmi iterativi che non esplorano l'intero spazio delle possibili partizioni ma solo una parte di esso: non v'è perciò garanzia di ottenere la soluzione ottima in senso assoluto.

Indice

Metodo delle k-medie

Consiste di un procedimento iterativo in cui si parte da una suddivisione arbitraria in g costituenti non vuoti

1. si calcolano i centroidi dei gruppi così definiti, dove il centroide è il vettore delle medie;
2. si alloca ogni osservazione al centroide più vicino;
3. si torna al passo (1) con la nuova partizione sino a che l'operazione di cui a (2) non modifica più la partizione.

Il risultato dipende dalla metrica utilizzata, se si utilizza quella euclidea è garantita la convergenza dell'algoritmo e la funzione obbiettivo in uso è inversamente proporzionale alla varianza tra gruppi e direttamente a quella interna ai gruppi.

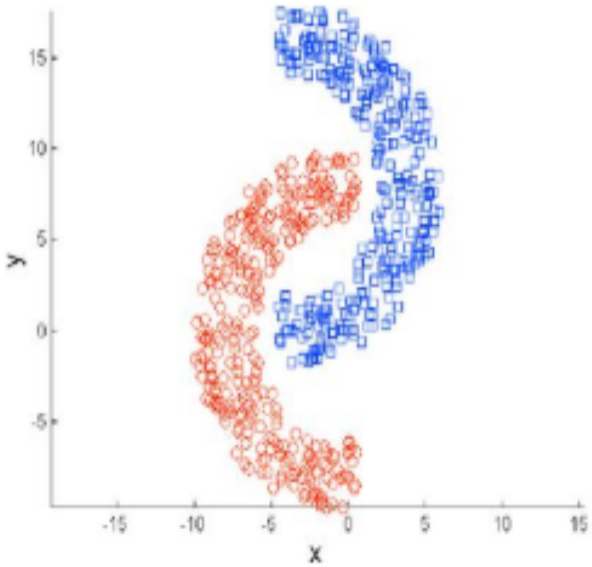
Metodo delle k-medie (2)

Vantaggi e svantaggi:

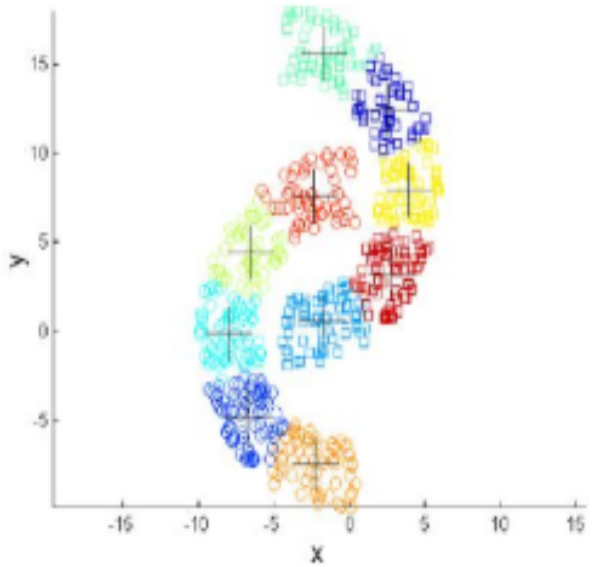
+ è relativamente efficiente;
+ adatto a scoprire gruppi di forma convessa;
- è inadatto per gruppi di forma concava;
- il metodo si applica al caso di dati continui;
- non è garantita la convergenza all'ottimo assoluto
- non è garantito che converga allo stesso punto partendo da punti di partenza (partizione iniziale) diversi;
- il risultato è sensibile alla presenza di valori anomali.

Esempio

Un esmpio di quando il metodo delle K-medie fallisce:



Gruppi reali



K-means Clusters

Metodo delle k-medie, esempio

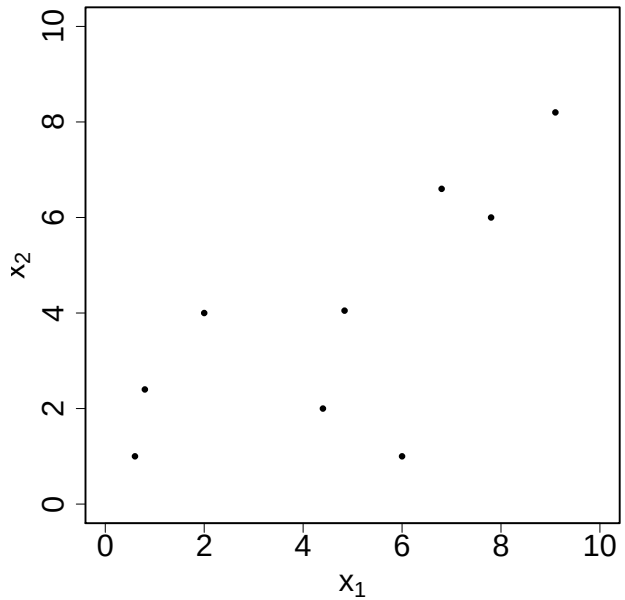
Consideriamo, a titolo di esempio, il campione (x_{i1}, x_{i2}) riportato nelle prime tre colonne della tabella

i	x_{i1}	x_{i2}	$g_i^{(0)}$	$d_{i0}^{(1)}$	$d_{i1}^{(1)}$	$g_i^{(1)}$	$d_{i0}^{(2)}$	$d_{i1}^{(2)}$	$g_i^{(2)}$	$d_{i0}^{(3)}$	$d_{i1}^{(3)}$	$g_i^{(3)}$
1	0.8	2.4	1	5.54	3.24	1	7.39	1.99	1	8.42	2.31	1
2	2.0	4.0	0	3.77	2.29	1	5.59	2.06	1	6.59	1.94	1
3	7.8	6.0	1	2.36	4.90	0	0.70	6.38	0	0.94	5.91	0
4	4.4	2.0	0	3.41	0.97	1	5.02	1.64	1	6.05	1.36	1
5	6.0	1.0	1	4.22	2.75	1	5.33	3.42	1	6.23	3.22	1
6	9.1	8.2	0	4.63	7.36	0	2.80	8.81	0	1.74	8.33	0
7	0.6	1.0	1	6.51	3.90	1	8.36	2.41	1	9.41	2.87	1
8	6.8	6.6	0	1.86	4.64	0	0.51	6.06	0	1.15	5.59	0
9	4.8	4.0	1	1.37	1.43	0	3.15	2.86	1	4.21	2.39	1

Metodo delle k-medie, esempio (2)

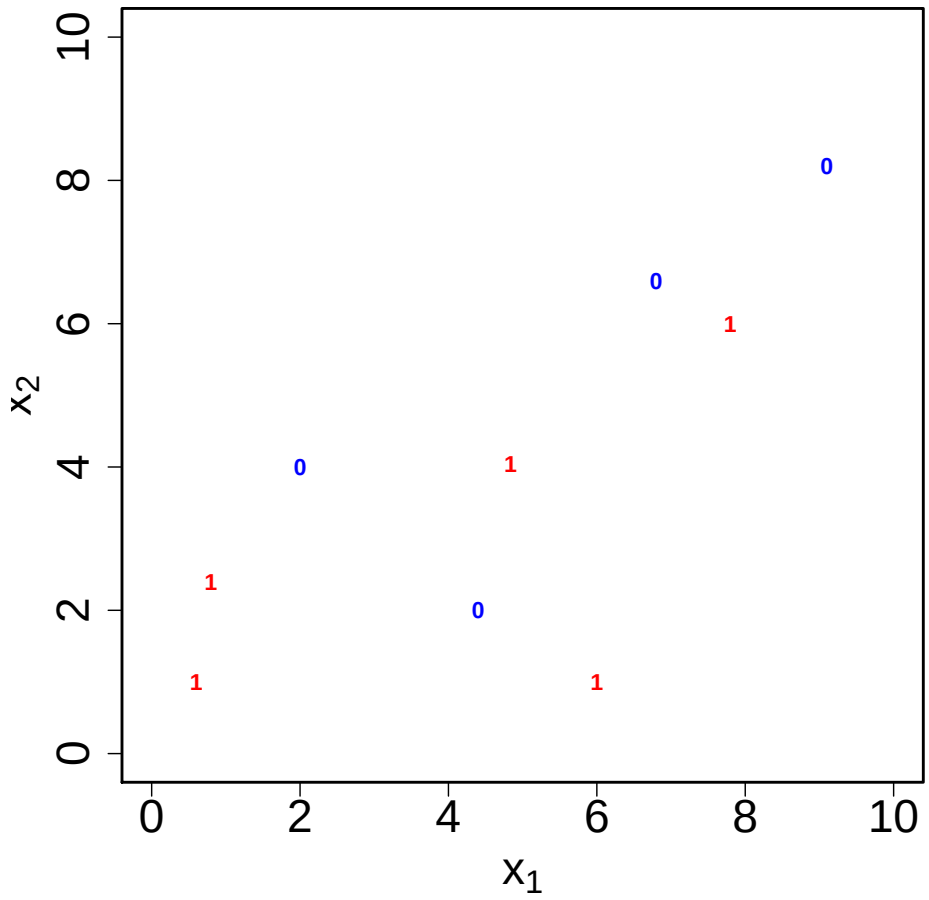
Assumiamo

- le variabili siano tali per cui è legittimo utilizzare la distanza euclidea tra osservazioni
- $g = 2$;



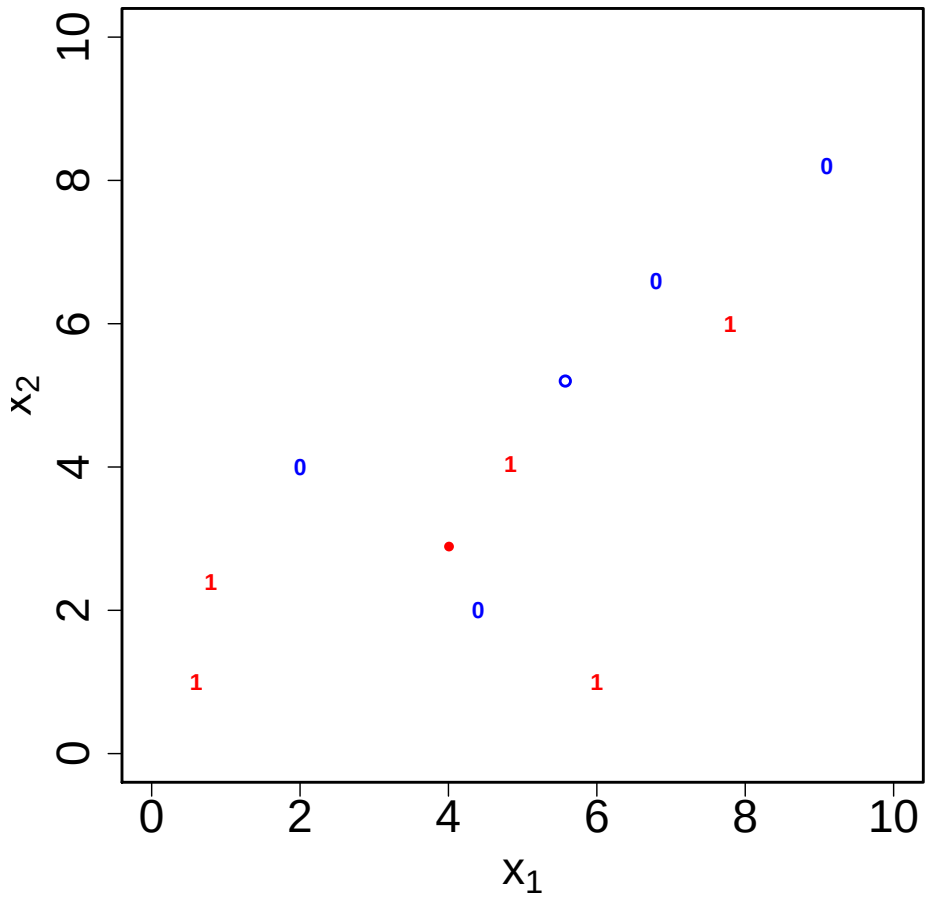
Metodo delle k-medie, esempio (3)

Al primo passo attribuiamo dei gruppi in modo casuale



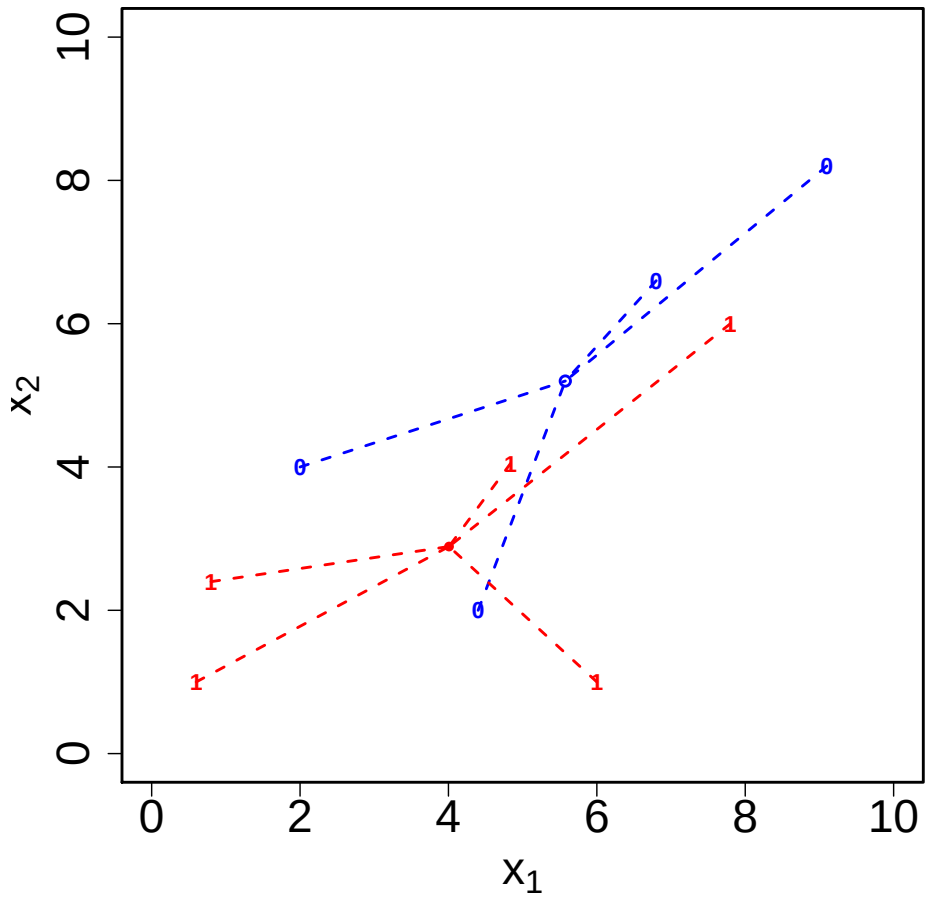
Metodo delle k-medie, esempio (4)

Calcoliamo i centroidi



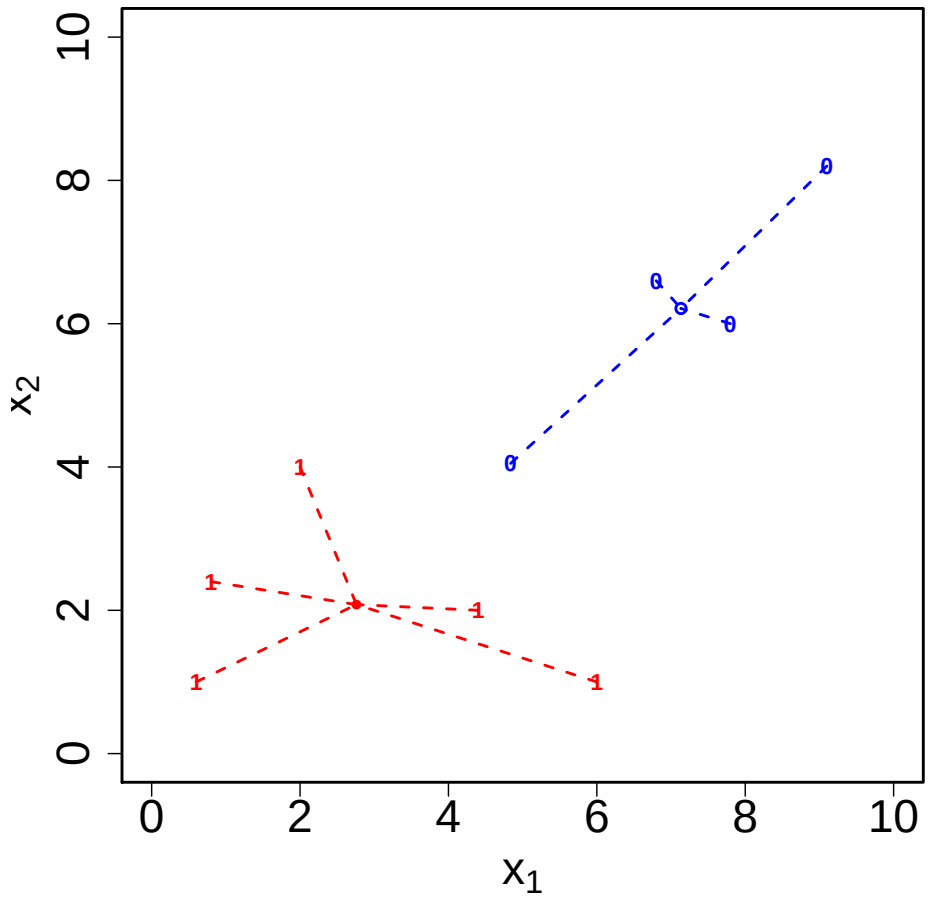
Metodo delle k-medie, esempio (5)

e quindi le distanze

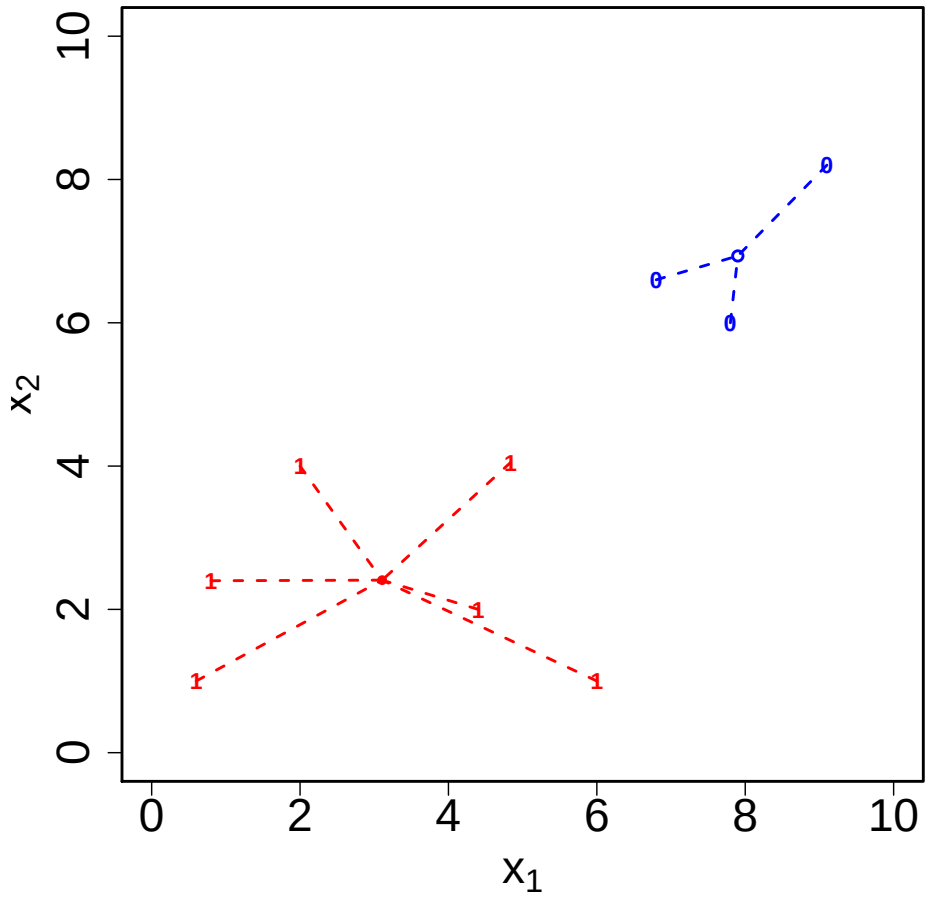


Metodo delle k-medie, esempio (6)

- Si riassegnano le osservazioni e si ripete la procedura

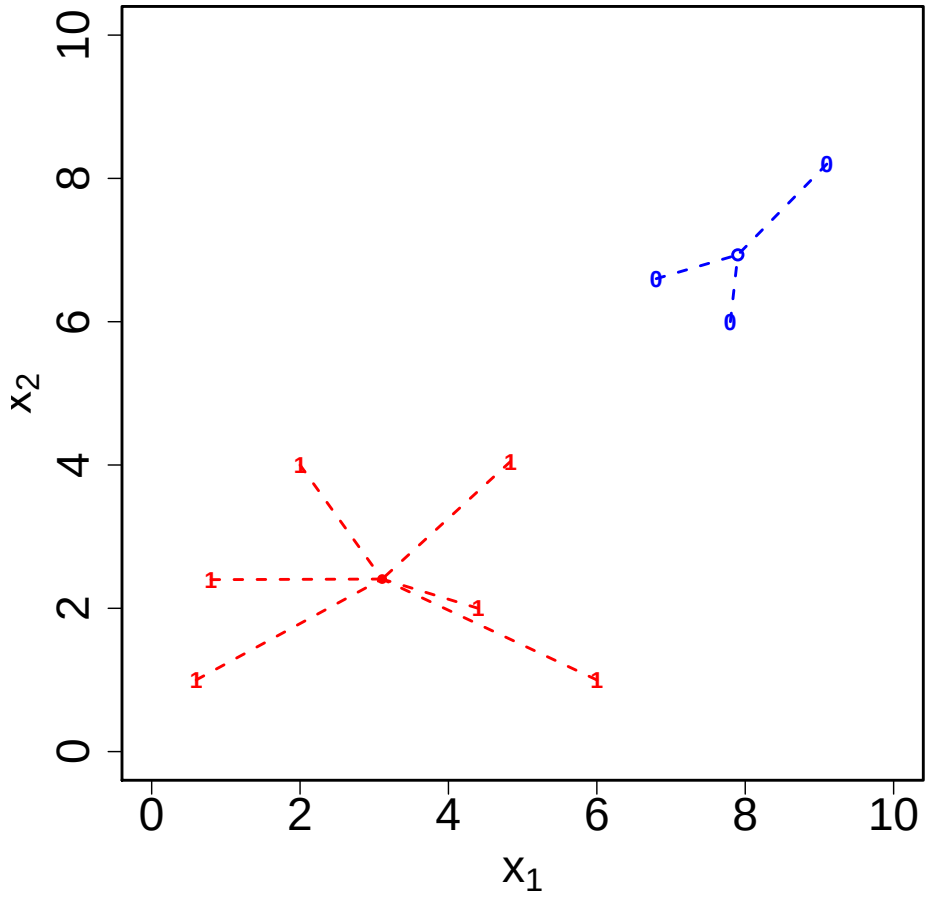


Metodo delle k-medie, esempio (7)



Metodo delle k-medie, esempio (8)

- Un'ulteriore iterazione non porta a modifiche dei gruppi



Indice

Metodi gerarchici

Si prescinde dalla scelta della numerosità della partizione, ma si crea una famiglia di partizioni in cui il numero dei gruppi varia da 1 a n in modo che la partizione in $i+1$ gruppi sia ottenuta dalla partizione in i facendo di uno degli elementi di questa due elementi di quella.
Tale insieme di partizioni può essere ottenuto o per successive agglomerazioni, si parla allora di algoritmo agglomerativo (AGNES) o per successivi raffinamenti delle partizioni, si parla allora di algoritmo divisivo (DIANA).

Metodi gerarchici

Nel caso di algoritmo agglomerativo

1. si parte dalla partizione in n gruppi ciascuno singoletto;
2. si determina tra le $\binom{n}{2} = n(n-1)/2$ coppie di gruppi quale sia quella migliore da unire;
3. si forma una nuova partizione, in $n-1$ costituenti unendo i due elementi selezionati.

Iterando questo procedimento si perviene in $n-1$ passi alla partizione in un unico gruppo (ovviamente, il campione intero).
Dobbiamo precisare come si sceglie la coppia da unire. Vi sono diverse opzioni, nessuna nettamente superiore a un'altra e la scelta andrà fatta caso a caso.
Fatto sta che vanno uniti quei gruppi che al passo precedente risultano i **più vicini** confrontando tutte le possibili coppie di gruppi

Distanze tra gruppi

Possiamo definire una distanza tra gruppi, con cui calcolare le distanze tra le $n(n-1)/2$ coppie (di gruppi) e scegliere come da unire la coppia con gli elementi più vicini.
Sia $d(\cdot,\cdot)$ la distanza tra unità e consideriamo due gruppi R_1 e R_2, possibili definizioni di distanze tra i gruppi sono

- **metodo del legame singolo** (*single linkage*) o metodo del vicino più vicino (***nearest neighbour***)

$$\delta(R_1, R_2) = \min\{d(\mathbf{x}_i, \mathbf{x}_j) : \mathbf{x}_i \in R_1, \mathbf{x}_j \in R_2\},$$

- **metodo del legame completo** (*complete linkage*) o vicino più lontano (*farthest neighbour*)

$$\delta(R_1, R_2) = \max\{d(\mathbf{x}_i, \mathbf{x}_j) : \mathbf{x}_i \in R_1, \mathbf{x}_j \in R_2\}$$

Distanze tra gruppi (2)

- **metodo della distanza media**

$$\delta(R_1, R_2) = \frac{1}{n_1 n_2} \sum_{\mathbf{x}_i \in R_1} \sum_{\mathbf{x}_j \in R_2} d(\mathbf{x}_i, \mathbf{x}_j)$$

dove n_1 e n_2 sono le cardinalità di R_1 e R_2 rispettivamente

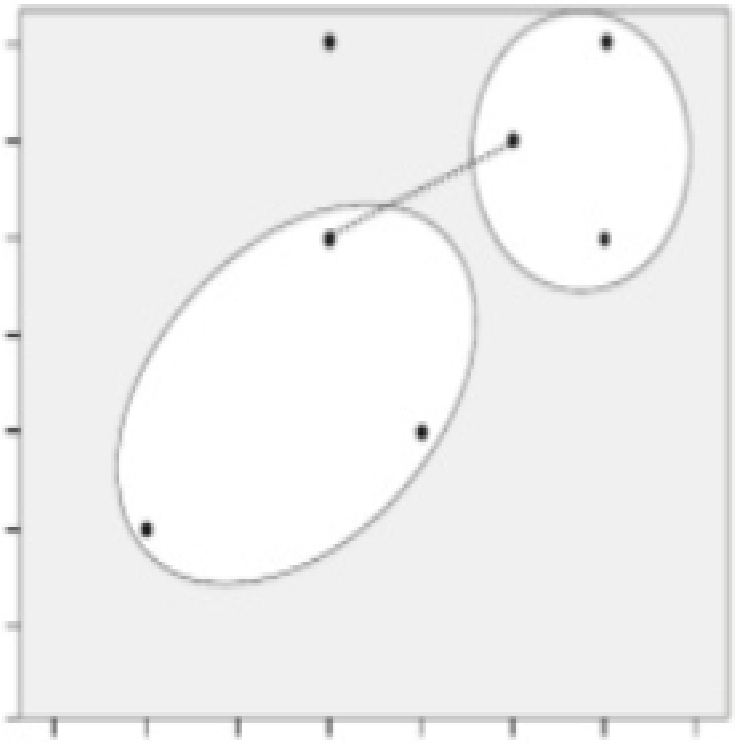
- **metodo della distanza tra le medie** (distanza cioè dei centroidi)

$$\delta(R_1, R_2) = \frac{1}{n_1 n_2} d(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2)$$

dove $\bar{\mathbf{x}}_j^T = \left(\sum_{\mathbf{x}_i \in R_j} x_{i1}, \ldots, \sum_{\mathbf{x}_i \in R_j} x_{in_j} \right)$ è il vettore delle medie delle variabili calcolato nel gruppo j.

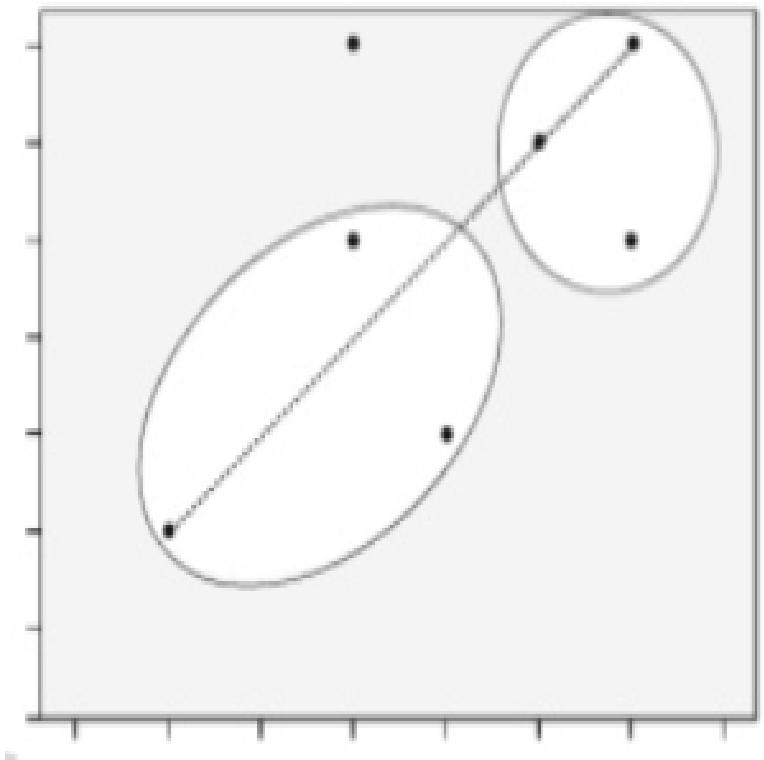
Distanze tra gruppi (3)

Metodo del legame singolo (nearest neighbor)



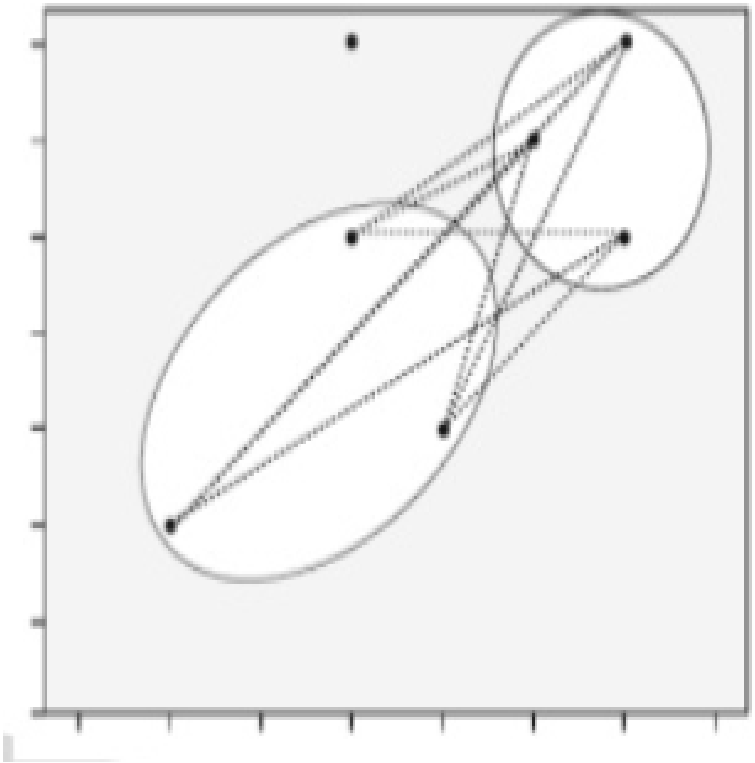
Distanze tra gruppi (4)

Metodo del legame completo



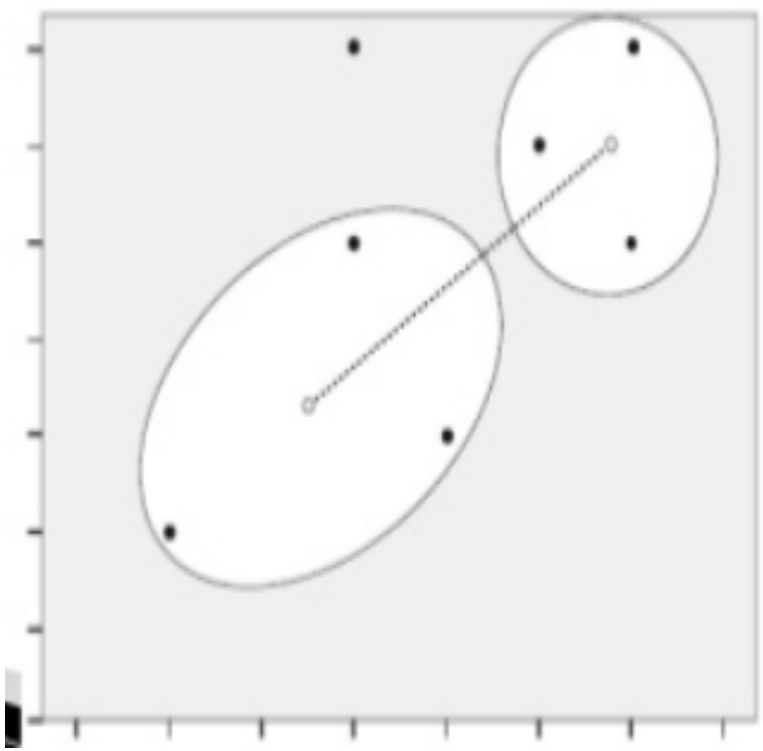
Distanze tra gruppi (5)

Metodo del legame medio



Distanze tra gruppi (6)

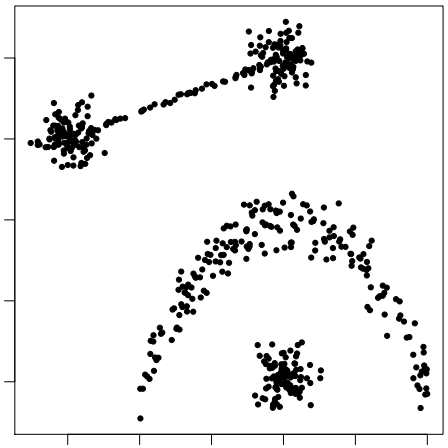
Metodo del centroide



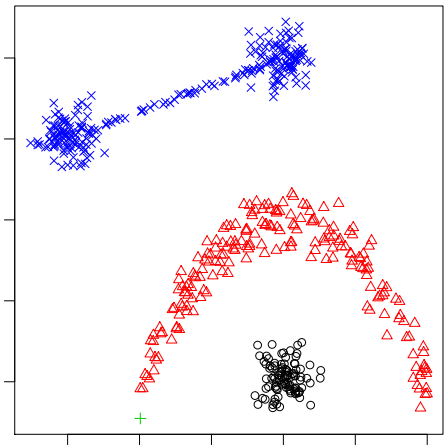
Distanze tra gruppi (7)

Una peculiarità del metodo del legame singolo è l'effetto catena, che da un lato consente di cogliere gruppi di forma particolare, dall'altro rischia di legare osservazioni che non appartengono a uno stesso gruppo.
Il metodo del legame completo, d'altra parte, tende a individuare gruppi molto compatti al loro interno ma di forma circolare (ipersferica, in generale) quindi si rischia di perdere gruppi di forma irregolare.

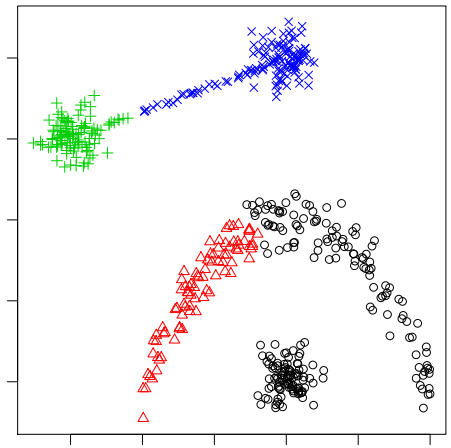
dati



singolo



completo



Criterio di Ward

Si basa il confronto tra le coppie di gruppi sulle varianze intorno alle medie di gruppo.
Alla partizione in G gruppi $\{R_1, \ldots, R_G\}$ si associa la varianza (errore)

$$\sum_{g=1}^{G} \sum_{\mathbf{x}_i \in R_g} d(\mathbf{x}_i, \bar{x}^{(g)})^2$$

dove $\bar{x}^{(g)}$ è la media del gruppo (vettore delle medie) e d è la distanza euclidea.
Se si uniscono due gruppi, g_1 e g_2 l'errore diventa

$$\sum_{g \neq g_1, g_2} \sum_{\mathbf{x}_i \in R_g} d(\mathbf{x}_i, \bar{x}^{(g)})^2 + \sum_{g \in \{g_1, g_2\}} \sum_{\mathbf{x}_i \in R_g} d(\mathbf{x}_i, (\bar{x}^{(g_1)} + \bar{x}^{(g_2)})/2)^2$$

Si calcola dunque la quantità sopra per tutte le coppie e si unisce la coppia alla quale corrisponde errore inferiore.

Esempio

Si abbia, per 6 osservazioni, la matrice di distanza

	1	2	3	4	5	6
1	0.0	11.0	9.0	11.1	22.3	15.1
2	11.0	0.0	11.6	17.5	18.7	16.3
3	9.0	11.6	0.0	9.8	14.7	7.4
4	11.1	17.5	9.8	0.0	22.7	13.8
5	22.3	18.7	14.7	22.7	0.0	10.6
6	15.1	16.3	7.4	13.8	10.6	0.0

Consideriamo un algoritmo agglomerativo in cui la distanza sia calcolata col metodo del legame singolo.

Esempio (2)

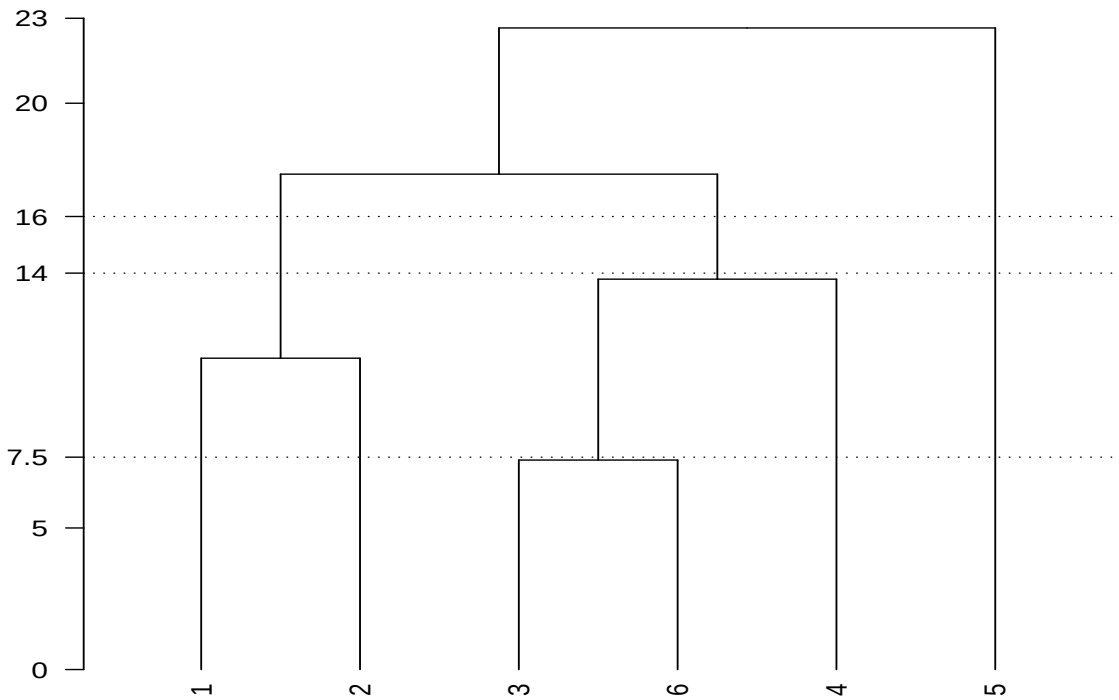
1. si raggruppano le unità 3 e 6, che hanno distanza 7.4, la nuova partizione è $\{\{1\},\{2\},\{4\},\{5\},\{3,6\}\}$ e la nuova matrice di distanza è

	1	2	4	5	3,6
1					
2	11.0				
4	11.1	17.5			
5	22.3	18.7	22.7		
3,6	9.0	11.6	9.8	10.6	

2. si fondono i gruppi $\{1\}$ e $\{3,6\}$,
3. ...

Il dendrogramma

È un metodo di rappresentazione di una gerarchia di partizioni, che evidenzia i gruppi che si formano a ogni stadio della classificazione



Il dendrogramma (2)

L'altezza del segmento che unisce due unità è generalmente corrispondente alla distanza tra essi (i gruppi singoletti $\{3\}$ e $\{6\}$ hanno distanza pari a 7.5 (circa)) (generalmente perché in qualche caso, per garantire leggibilità al grafico, è possibile che le si disegni a distanze uniformi).
Fissata una distanza, ad es. 10, si legge sul dendrogramma quali sono i gruppi che hanno distanza tra loro eguale o superiore a 10 disegnando una linea orizzontale ad altezza 10 e notando quali osservazioni vengono unite sino a quel punto, nell'esempio risultano formati i gruppi $\{1, 3, 4, 6\}$, $\{2\}$ e $\{5\}$.

Esempio

Si abbia, per 6 osservazioni, la matrice di distanza

	1	2	3	4	5	6
1	0.0	11.0	9.0	11.1	22.3	15.1
2	11.0	0.0	11.6	17.5	18.7	16.3
3	9.0	11.6	0.0	9.8	14.7	7.4
4	11.1	17.5	9.8	0.0	22.7	13.8
5	22.3	18.7	14.7	22.7	0.0	10.6
6	15.1	16.3	7.4	13.8	10.6	0.0

Consideriamo un algoritmo agglomerativo in cui la distanza sia calcolata col metodo del legame singolo.

Esempio (2)

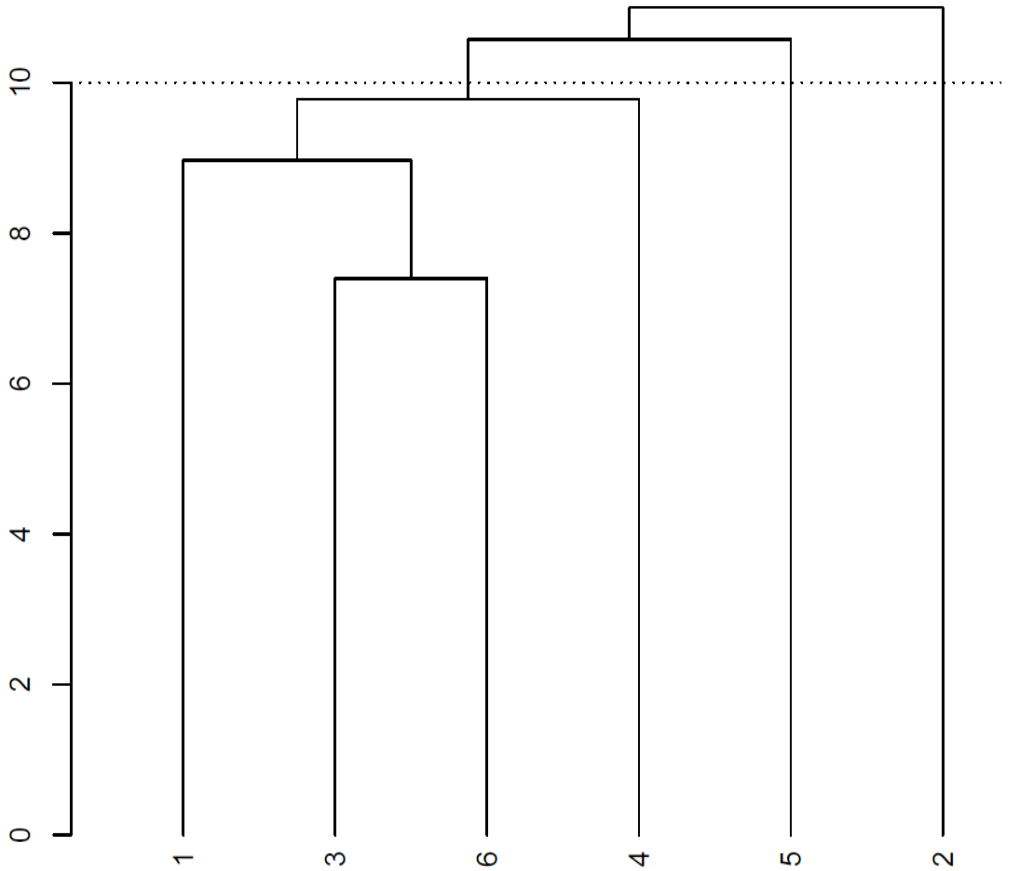
1. si raggruppano le unità 3 e 6, che hanno distanza 7.4, la nuova partizione è $\{\{1\},\{2\},\{4\},\{5\},\{3,6\}\}$ e la nuova matrice di distanza è

	1	2	4	5	3,6
1					
2	11.0				
4	11.1	17.5			
5	22.3	18.7	22.7		
3,6	9.0	11.6	9.8	10.6	

2. si fondono i gruppi $\{1\}$ e $\{3,6\}$,
3. ...

Il dendrogramma

È un metodo di rappresentazione di una gerarchia di partizioni, che evidenzia i gruppi che si formano a ogni stadio della classificazione



Il dendrogramma (2)

L' altezza del segmento che unisce due unità è generalmente corrispondente alla distanza tra essi (i gruppi singoletti $\{3\}$ e $\{6\}$ hanno distanza pari a 7.5 (circa)) (generalmente perché in qualche caso, per garantire leggibilità al grafico, è possibile che le si disegni a distanze uniformi).
Fissata una distanza, ad es. 10, si legge sul dendrogramma quali sono i gruppi che hanno distanza tra loro eguale o superiore a 10 disegnando una linea orizzontale ad altezza 10 e notando quali osservazioni vengono unite sino a quel punto, nell'esempio risultano formati i gruppi $\{1, 3, 4, 6\}$, $\{2\}$ e $\{5\}$.

Forza lavoro nei paesi europei

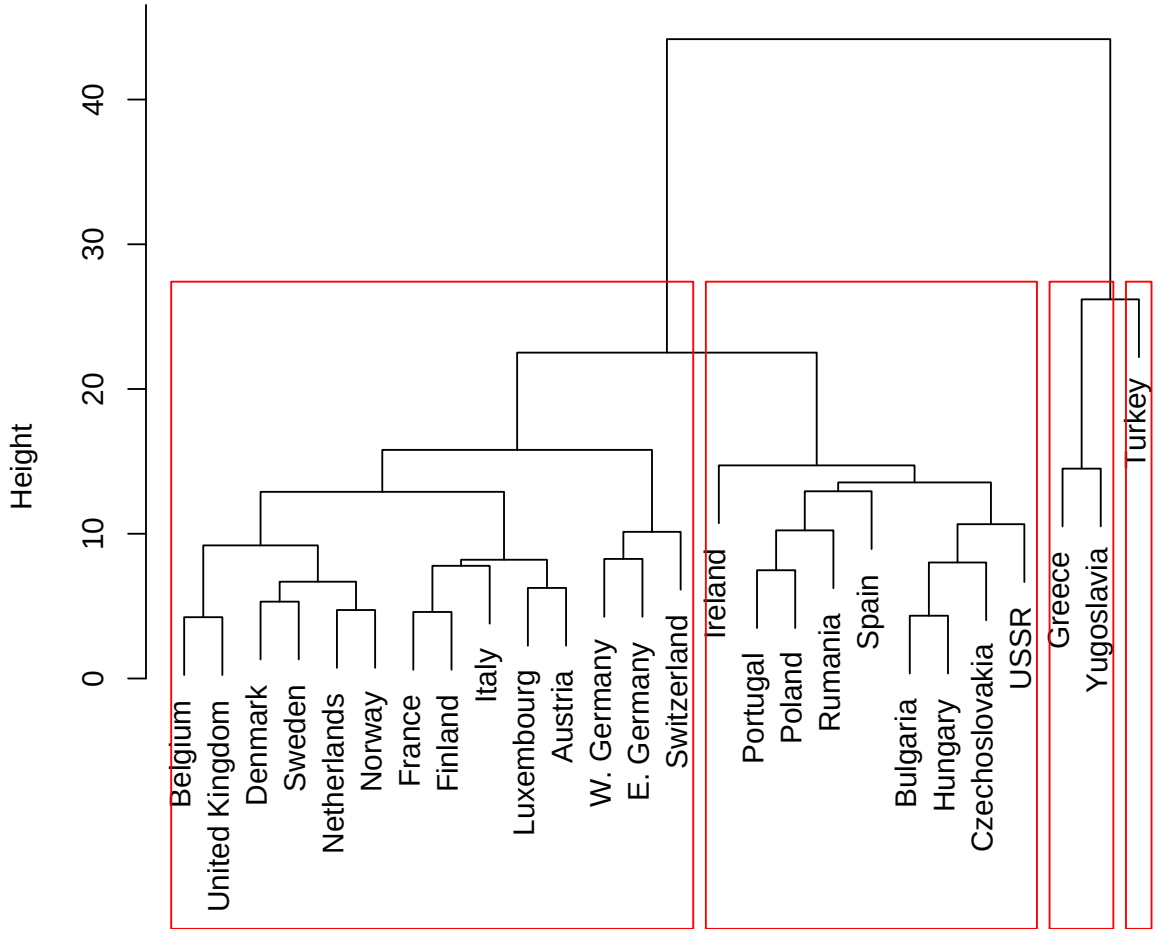
Si considerano le composizioni della forza lavoro per settore produttivo negli stati europei nel 1970,

Country	Agr	Min	Man	Pow	Con	Ser	Fin	SPS	TC	blocco
Belgium	3.3	0.9	27.6	0.9	8.2	19.1	6.2	26.6	7.2	w
Denmark	9.2	0.1	21.8	0.6	8.3	14.6	6.5	32.2	7.1	w
France	10.8	0.8	27.5	0.9	8.9	16.8	6.0	22.6	5.7	w
W. Germany	6.7	1.3	35.8	0.9	7.3	14.4	5.0	22.3	6.1	w
Ireland	23.2	1.0	20.7	1.3	7.5	16.8	2.8	20.8	6.1	w
Italy	15.9	0.6	27.6	0.5	10.0	18.1	1.6	20.1	5.7	w
Luxembourg	7.7	3.1	30.8	0.8	9.2	18.5	4.6	19.2	6.2	w
Netherlands	6.3	0.1	22.5	1.0	9.9	18.0	6.8	28.5	6.8	w
United Kingdom	2.7	1.4	30.2	1.4	6.9	16.9	5.7	28.3	6.4	w
Austria	12.7	1.1	30.2	1.4	9.0	16.8	4.9	16.8	7.0	w
Finland	13.0	0.4	25.9	1.3	7.4	14.7	5.5	24.3	7.6	w
Greece	41.4	0.6	17.6	0.6	8.1	11.5	2.4	11.0	6.7	w
Norway	9.0	0.5	22.4	0.8	8.6	16.9	4.7	27.6	9.4	w
Portugal	27.8	0.3	24.5	0.6	8.4	13.3	2.7	16.7	5.7	w
Spain	22.9	0.8	28.5	0.7	11.5	9.7	8.5	11.8	5.5	w
Sweden	6.1	0.4	25.9	0.8	7.2	14.4	6.0	32.4	6.8	w
Switzerland	7.7	0.2	37.8	0.8	9.5	17.5	5.3	15.4	5.7	n
Turkey	66.8	0.7	7.9	0.1	2.8	5.2	1.1	11.9	3.2	n
Bulgaria	23.6	1.9	32.3	0.6	7.9	8.0	0.7	18.2	6.7	e
Czechoslovakia	16.5	2.9	35.5	1.2	8.7	9.2	0.9	17.9	7.0	e
E. Germany	4.2	2.9	41.2	1.3	7.6	11.2	1.2	22.1	8.4	e
Hungary	21.7	3.1	29.6	1.9	8.2	9.4	0.9	17.2	8.0	e
Poland	31.1	2.5	25.7	0.9	8.4	7.5	0.9	16.1	6.9	e
Rumania	34.7	2.1	30.1	0.6	8.7	5.9	1.3	11.7	5.0	e
USSR	23.7	1.4	25.8	0.6	9.2	6.1	0.5	23.6	9.3	e
Yugoslavia	48.7	1.5	16.8	1.1	4.9	6.4	11.3	5.3	4.0	n

Forza lavoro nei paesi europei (2)

Applichiamo ad essi un metodo di raggruppamento gerarchico agglomerativo per confrontare i gruppi ottenuti con le caratteristiche del sistema economico (e in particolare con l'appartenenza al blocco occidentale o orientale).
Considerando la partizione in quattro gruppi, i gruppi formatisi comprendono uno prevalentemente i paesi del blocco occidentale, uno prevalentemente i paesi del blocco orientale, i due gruppi residuali comprendono Grecia e Yugoslavia uno e Turchia l'altro.
In questo caso, dunque, è possibile affermare che il raggruppamento ottenuto statisticamente ha una corrispondenza fattuale.
Scendendo nella gerarchia si nota come si formi presto un gruppo contenente i paesi del Nord-Europa (con l'esclusione della Finlandia).

Forza lavoro nei paesi europei (3)



Indice

Silhouette

Determinato, in qualunque modo, un raggruppamento di n unità in G gruppi $\{R_1, \ldots, R_G\}$ la *silhouette* della partizione è uno strumento per verificare la 'bontà' di tale partizione, ossia in che misura si abbia 'coesione interna e separazione esterna'.
Si confronta, per ciascuna osservazione, quanto essa sia vicina al suo gruppo e agli altri.
Sia d la distanza di riferimento, la distanza dell'osservazione $\mathbf{x}_{i^*}$ dal gruppo R_g è

$$D(\mathbf{x}_{i^*}, R_g) = \frac{1}{n_g} \sum_{\mathbf{x}_i \in R_g} d(\mathbf{x}_{i^*}, \mathbf{x}_i).$$

Sia poi R_{g^*} il gruppo in cui è inclusa l'osservazione $\mathbf{x}_{i^*}$ e sia

$$D_0 = \min_{g \neq g^*} D(\mathbf{x}_{i^*}, R_g),$$

Silhouette (2)

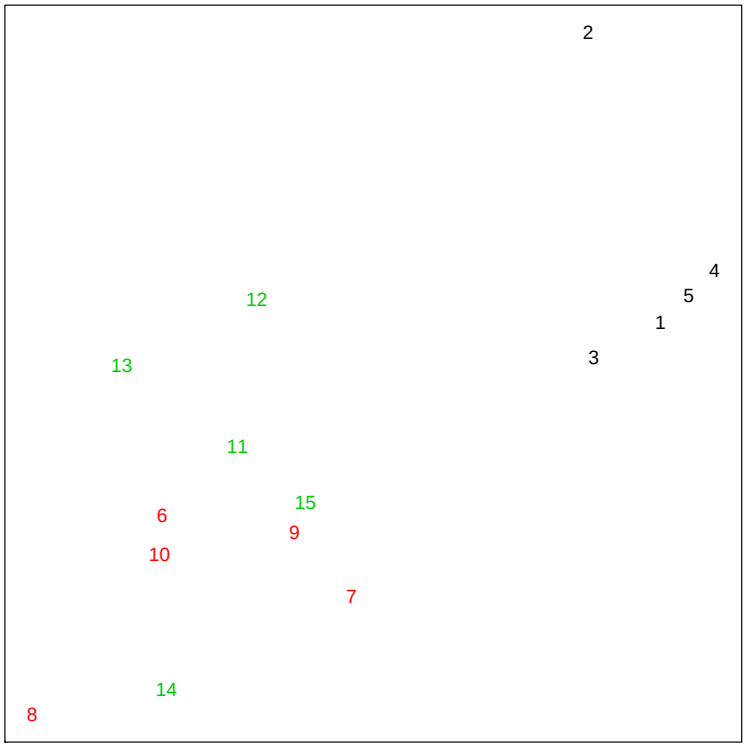
D_0 è la distanza di $\mathbf{x}_{i^*}$ dal gruppo più vicino diverso da quello cui appartiene, si confronta D_0 con la distanza dal suo gruppo mediante

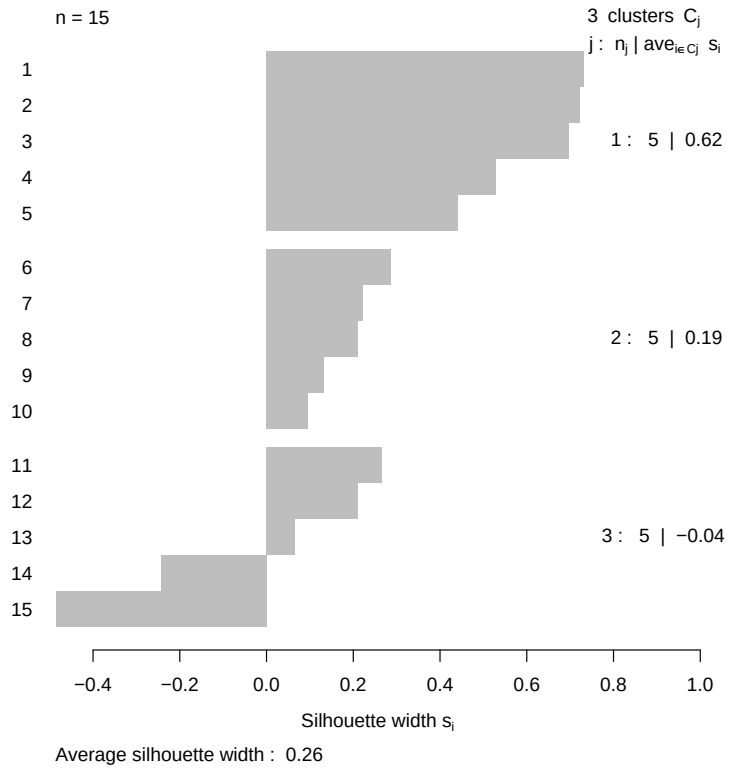
$$S(\mathbf{x}_{i^*}) = \frac{D_0 - D(\mathbf{x}_{i^*}, R_{g^*})}{\max\{D_0, D(\mathbf{x}_{i^*}, R_{g^*})\}}.$$

- $S(\mathbf{x}_{i^*}) \leq 1$
- $S(\mathbf{x}_{i^*})$ è tanto più grande quanto più $\mathbf{x}_{i^*}$ è vicino al suo gruppo e distante dagli altri gruppi.
- $S(\mathbf{x}_{i^*}) < 0$ indica che $\mathbf{x}_{i^*}$ è più vicino a un altro gruppo che non al suo.

I valori così ottenuti possono poi essere messi in un grafico che mostra quanto bene i gruppi siano separati tra di loro e permette di individuare eventuali osservazioni problematiche.

Silhouette (3)





Silhouette (4)

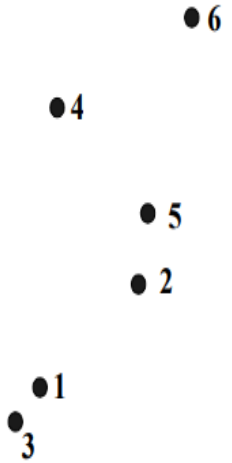
	cluster	neighbor	sil_width
hline a	1	3	0.69703258
b	1	3	0.44191866
c	1	3	0.52951658
d	1	3	0.72208882
e	1	3	0.73153062
f	2	3	0.13158934
g	2	3	0.20989431
h	2	3	0.22259232
i	2	3	0.09656938
j	2	3	0.28796815
k	3	2	0.06506903
l	3	2	0.26597938
m	3	2	0.21101039
n	3	2 -0.48648319	
o	3	2 -0.24282093	

Indice

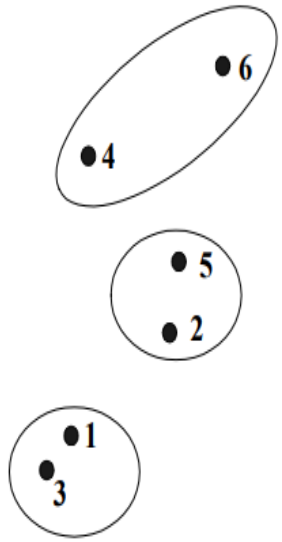
Confronti

Non è possibile determinare un criterio generale per la scelta di un metodo di raggruppamento dato un campione, il modo usuale di procedere consiste nell'applicare diversi metodi e confrontare i risultati ottenuti che, generalmente, saranno diversi anche in misura significativa: se però vi sono gruppi di osservazioni che vengono trattate nello stesso modo (accorpate o divise) da diverse tecniche, ciò è indicativo di una loro somiglianza o diversità. Anche l'ausilio di tecniche grafiche può essere utile per capire o interpretare i raggruppamenti di osservazioni.

Confronti

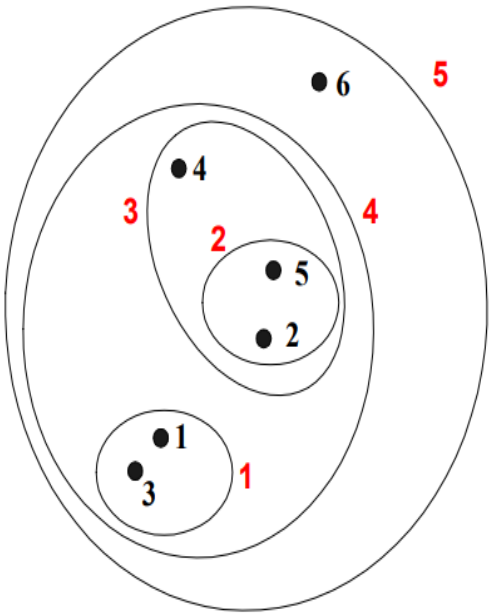


Punti originali

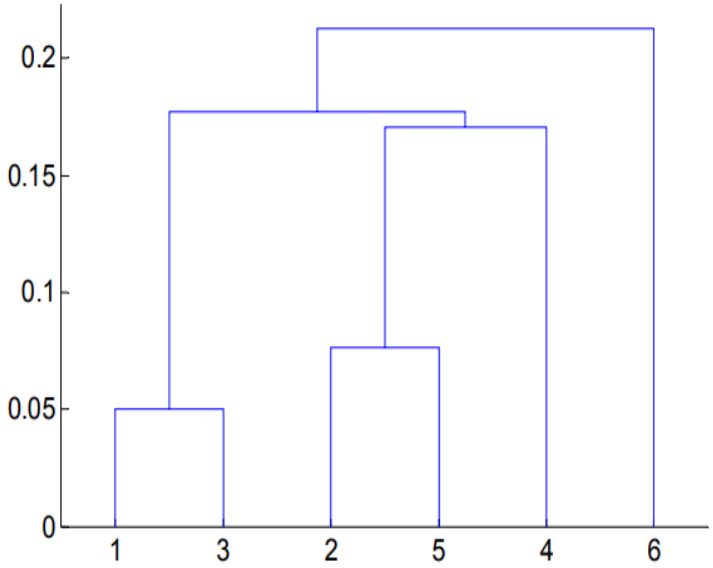


Partitional Clustering

Confronti



Traditional Hierarchical Clustering



Traditional Dendrogram