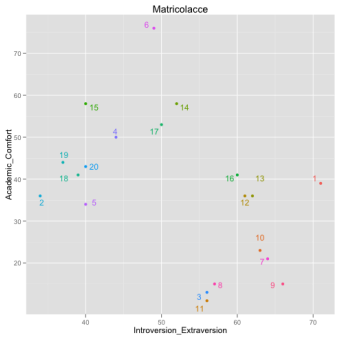
Esempio introduttivo

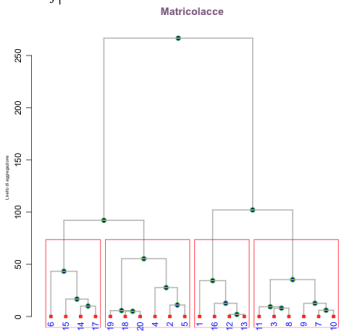
Un gruppo di 20 matricolacce è stato classificato in base a due punteggi relativi ad un test attitudinale ed un test comportamentale.



Si presuppone (ovvero si vuole verificare) che esistano due o più gruppi distinti di soggetti.

Si stabilisce una misura di distanza tra le unità (matricolacce). Ad esempio:

$$d_{ij} = \sum_{r=1}^{m} |x_{ri} - x_{rj}|, \; con \;\; m = 2 \;\; \text{variabili}$$



Si opta per un approccio. Ad esempio gerarchico agglomerativo- con link di Ward.

Sono plausibili due suddivisioni: in 2 o in 4 gruppi.

Analisi dei gruppi (cluster analysis)

Partiamo dal presupposto che entità diverse siano soggette a forze diverse che possono attivare dati con un impatto differenziato in relazione al gruppo a cui l'entità appartiene.

Si presuppone che le unità o le variabili (in generale, le entità) siano classificabili in raggruppamenti naturali, ovvero si vuole sottoporre a verifica l'ipotesi che fra le entità esistano dei sottoinsiemi distinti

Un primo obiettivo è di identificare un numero ridotto di profili caratteristici in base al contenuto informativo che si può ricavare dal data set.

Lo scopo evidente è quello poi di organizzare i valori in ragione della tipologia di entità per articolarne più compiutamente l'interpretazione.

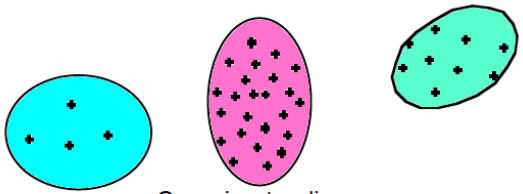
Analisi dei gruppi/2

L'idea è simile a quella che sta alla base della suddivisione in classi dei valori di una variabile, solo che ora avviene con dati multivariati.

Gli scopi sono anche gli stessi:

Semplificare individuando pochi casi tipici in alternativa alla totalità degli oggetti da trattare.

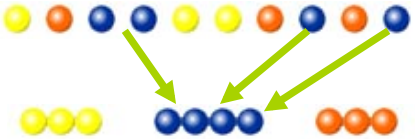
Si presuppone in effetti che le entità siano classificabili in gruppi naturali, ovvero si vuole sottoporre a verica l'ipotesi che fra le entità esistano dei sottogruppi distinti e distanti.



Gruppi naturali

Finalità della cluster analysis

Il raggruppamento o clustering può anche essere visto come una procedura che cerca divisioni interne plausibii di un data set ritenuto troppo grande per essere trattato come unico.



In questo la CA ricalca un'attività istintiva della mente umana di confrontare oggetti diversi cercandone somiglianze e differenze per categorie via via più distinte.

Non si conosce a priori né il numero né la struttura dei gruppi ne per quanto tempo e per quale numerosità una certa struttura rimane stabile.

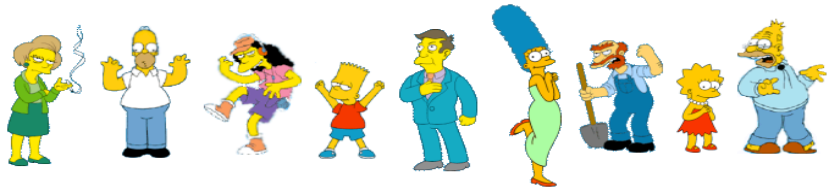
Finalità della cluster analysis/3

Extract Common Features:
- Sex
- Glasses
- Smile
- Hat
- Moustache

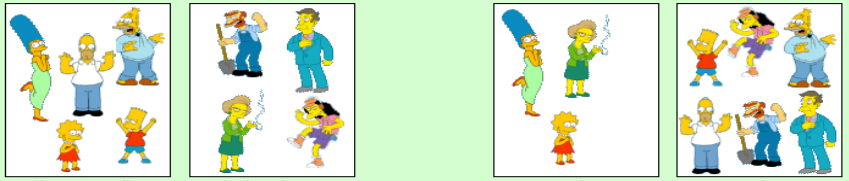
a) Al fine di chiarirsi le idee e meglio comunicarle (riduzione dei dati);
b) Per scoprire nuovi soggetti di ricerca (oulier o gruppi particolari);
c) Per pianificare una struttura organizzativa (zoning territoriale);
d) Per stabilire un elenco di riscontro (verifica di ipotesi);
e) Per definire particulari profili di entità;
f) Perché si hanno dei dati e qualcosa bisogna pur farne.

Cluster naturali

Kruskal (1977):" ... We call clusters natural if the membership is determined fairly well in a natural way by the data, and we call the clusters arbitrary if there is a substantial arbitrary element in the assignment process".



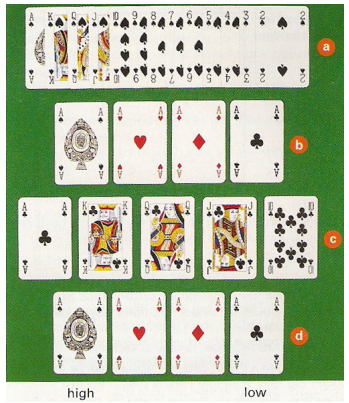
Clustering is subjective



Cluster naturali/2

Ad esempio le carte di un mazzo francese si raggruppano per seme se si gioca a Bridge, ma si raggruppano per valore se si gioca a Ramino.

Poiché i risultati sono influenzati sia dall'obiettivo della indagine che dal contesto applicativo, occorre effettuare delle scelte che individuino la procedura più adatta.

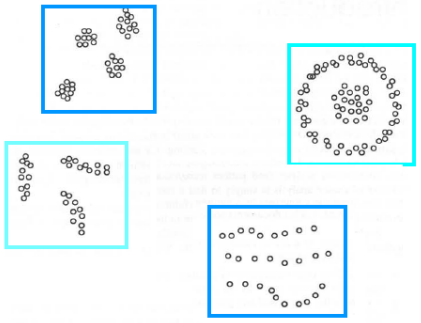


La scelta delle variabili da usare per la clustering determina il tipo di cluster che otteniamo

Cluster naturali/3

D.W. Goodall nel 1954 affermò:

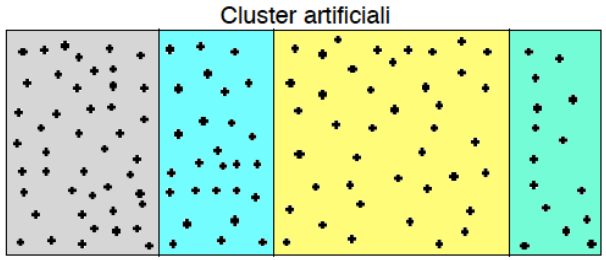
"A preference for classification is developed in childhood and persists as a habitual form of thought in adulhood"



A volte la natura dei gruppi è percepibile solo visivamente. Sarebbe difficile cogliere la struttura dei gruppi con tecniche solo numeriche

Cluster artificiali

Un insieme di "n" entità presenta dei cluster artificiali se l'appartenenza ai gruppi è soprattutto determinata da fattori esterni ai dati: ordine di considerazione, formule di calcolo, metodi e scale di misurazione.



Questo tipo di clustering è utile per determinare suddivisioni di unità amministrative in sezioni più piccole; per destinare senza una regola sistematica le entità a gruppi minori: casi per magistrati, esaminandi per sottocommissioni, etc.

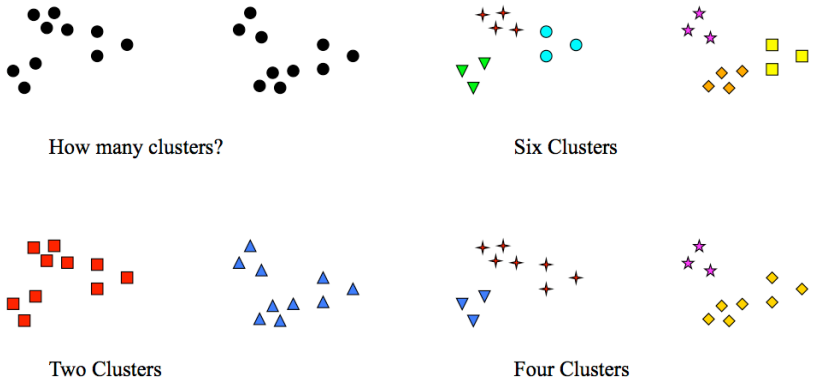
Perché è difficile

La definizione di gruppo è molto labile e, a seconda del tipo di analisi o del tipo di dati, si perviene a raggruppamenti diversi.

Si debbono quindi adottare definizioni intuitive, pragmatiche che danno regole operative, ma non propongono concetti nitidi.

Ad esempio è più facile spiegare come riconoscere un cluster che dire che cosa è un cluster.

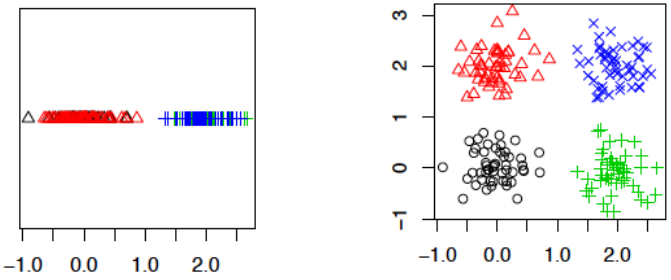
Perché è difficile/2



Sebbene non sia semplice immaginare una tecnica con basi così labili ed evanescenti è altrettanto complicato trovare una tecnica di analisi così tanto diffusa

Perché è difficile/3

La cluster analysis è una procedura multivariata. La scelta di quante e quali variabili adoperare è fondamentale.



I quattro raggruppamenti sono evidenti nel grafico bivariato a destra, ma si riesce a percepirne solo due nel grafico univariato a sinistra

Classificazione

Si discute di n entità $T=\{U_1,U_2,\ldots,U_n\}$ per le quali è nota la matrice delle dissimilarità o distanze

$$D=\begin{array}{c|cccc} & U_1 & U_2 & U_i & U_n \\ \hline U_1 & 0 & d_{12} & d_{1i} & d_{1n} \\ U_2 & d_{12} & 0 & d_{2i} & d_{2n} \\ U_i & d_{1i} & d_{2i} & 0 & d_{in} \\ U_n & d_{1n} & d_{2n} & d_{in} & 0 \end{array}$$

Ipotizziamo che le entità rientrino in k gruppi $C_1, C_2,\ldots, C_k$ tali che

$$C_r \cap C_s = \varnothing \quad per \quad r \neq s$$
$$\bigcup_{r=1,\cdots,k} C_r = T, \quad con \quad C_r \neq \varnothing, \; r = 1,\cdots,k$$

Ogni entità entra in uno ed un solo gruppo

Classificazione/2

Ad ogni entità è associato il valor dell'indicatore di gruppo

$$\gamma_j = r \quad se \quad e \quad solo \quad se \quad U_j \in C_r, \; j = 1,2,\cdots,n$$

Ogni cluster contiene un certo numero di entità (cardinalità del cluster)

$$n_r = \sum_{j=1}^{n} I\left(\gamma_j = r\right), \; n_r > 0, \; \sum_{r=1}^{k} n_r = n$$

Tutte le unità sono classificate. Ogni unità ricade in uno dei gruppi

La funzione I(x) è un indicatore di condizione logica tale che

$$I(x) = \begin{cases} 1 \; se \; x \; è \; vera \\ 0 \; se \; x \; è \; falsa \end{cases}$$

Riconoscimento dei gruppi

I gruppi esistono se sono ben differenziati in base a due principi confliggenti:

COESIONE INTERNA O COMPATEZZA (entità nello stesso cluster sono operativamente simili le une alle altre).

ISOLAMENTO ESTERNO O SEPARAZIONE (entità molto dissimili debbono collocarsi in cluster diversi).

I due principi sono connessi:
1) Cluster molto densi richiedono meno isolamento.
2) Cluster molto sparsi sono accettabili se grandemente separati dal resto.
3) L'omogeneità tende ad essere maggiore per i cluster piccoli. La separazione ha una tendenza opposta.

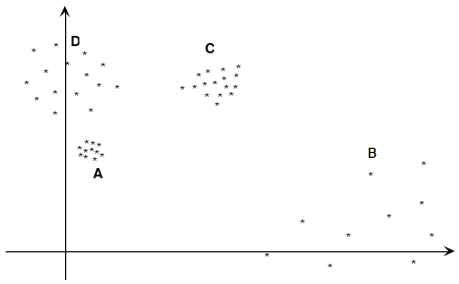
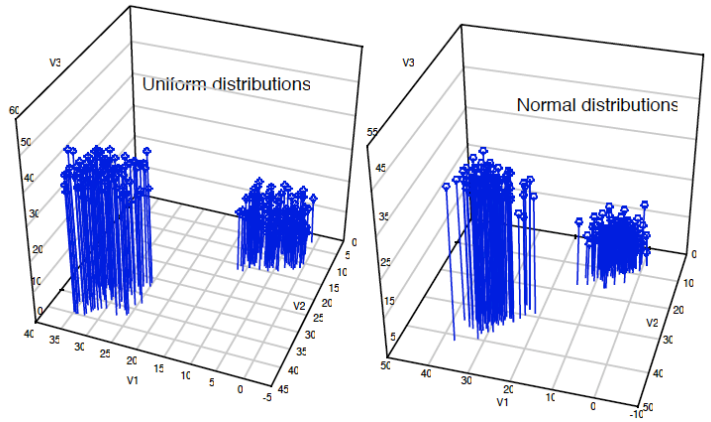


Figure 2: compactness and isolation of clusters

Riconoscimento dei gruppi/2

La situazione che si ipotizza nei dati è la seguente:



Nei dati reali questa però è un miraggio. Tuttavia, se gli algoritmi di classificazione non danno almeno qui dei buoni risultati, sono da scartare.

Classificazione ideale

Le tecniche per l'analisi dei gruppi hanno uno scopo comune ed intuitivo: entità collocate in uno stesso cluster hanno affinità/somiglianza maggiore rispetto ad entità collocate in gruppi diversi.

Raggruppamenti "ben strutturati". La distanza massima fra entità interne ad uno stesso cluster deve essere inferiore alla distanza minimmafra entità in cluster diversi,

Diametro del cluster

$$\Delta_r = \max_{\gamma_i = r, \gamma_j = r} \left\{ d_{i,j} \right\}, \; i,j = 1,2,\cdots,n;$$

Separatezza o isolamento del cluster

$$\Psi_{r,s} = \min_{\substack{\gamma_i = r, \gamma_j = s \\ r \neq s}} x \left\{ d_{i,j} \right\}, \; i,j = 1,2,\cdots,n$$

La classificazione è ideale se il diametro di ogni cluster è minore di ogni sua separatezza.

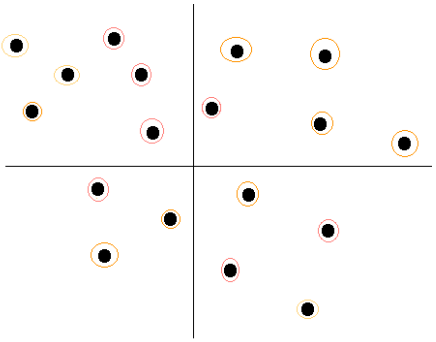
$$\Delta_r \leq \max_{\substack{s=1,\cdots,k \\ s \neq r}} \Psi_{r,s}$$

Esempio

Un esempio di classificazione ben strutturata è la clustering

$$\gamma_j = j, \; j = 1,2,\cdots,n$$

In cui ogni entità forma un gruppo autonomo



Non è molto utile dato che così non è stata effettuata alcuna sintesi delle informazioni

Lo scopo della cluster analysis è di sintetizzare al meglio le informazioni anche a costo di disperderne una parte cospicua.

Altro esempio

Un gruppo deve essere omogeneo al suo interno, ma eterogeneo rispetti agli altri gruppi

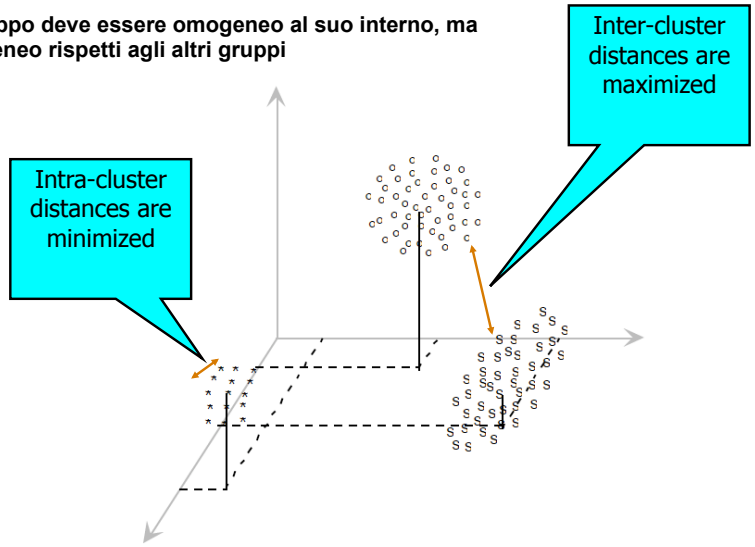


Figure 1: clusters of different shapes and sizes

Cluster ibridi

Il cluster G si colloca in un processo che trasforma le unità in C1 in unità di C3 e/o viceversa

Il cluster H è una struttura che è formata dalle entità che condividono parte delle peculiarità di C1 ed in parte quelle di C2.

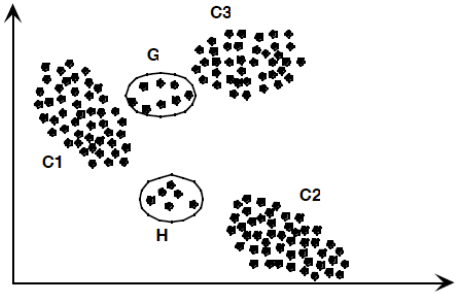


Figure 3a: hybrid clusters

Se si chiede all'algoritmo di clustering di travasare le unità in H o in G in uno dei cluster più prossimi rimarrà comunque il dubbio sul loro effettivo isolamento e sulla loro compattezza.

Cannibalizzazione e Singoletti

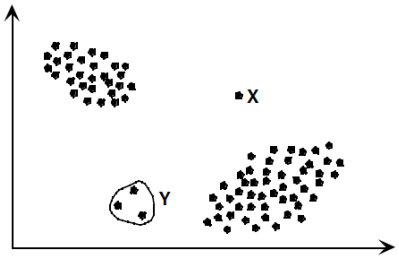


Figure 3b: a singleton and a small cluster

Il cluster Y contiene sufficienti unità da poter essere considerato effettimante esistente nella matrice dei dati oppure è un mero artefatto delle procedure per il trattamento dei dati?
Non esiste una risposta certa.

Il cluster X formato da una sola unità è un singoletto. Se è sufficientemente distante dagli altri gruppi occorre considerarlo anche se è fonte di problemi per la descrizione del cluster (ad esempio non c'è variabilità).

Se X è prossimo ad un altro cluster si può pensare ad una unità anomala da rielaborare.

Sia X che Y rendono difficile la scelta del numero di cluster: 3 (ignorando X) o 4 (considerando X un gruppo autentico, per quanto ridotto)

Le fasi della cluster analysis

Le fasi del processo di analisi dei gruppi si concretizzano in una serie di decisioni da prendere in merito a diverse scelte.

In particolare:

a) scelta delle entità di analisi;

b) scelta delle variabili caratterizzanti ciascuna entità;

c) omogeneizzazione delle scale di misura utilizzate per esprimere le diverse caratteristiche considerate;

d) scelta della misura di dissimilarità o di distanza tra le entità;

e) definizione del numero di gruppi che si vogliono o di debbono formare;

f) scelta dell'algoritmo di classificazione;

g) interpretazione dei risultati ottenuti;

h) Validazione dei risultati

Classificazione delle procedure/2

ESPLORATIVE *(structure-seeking)*

INVASIVE *(structure-imposing)*

Laddove non esiste una struttura di gruppo, il metodo stesso provvederà ad imporne una, buona o cattiva che sia.

Le procedure di clustering operano sempre in maniera tale da collocare tutti gli elementi nei vari gruppi, indipendentemente dalla voglia o dall'idoneità delle varie unità a far parte di un gruppo in particolare.

Inoltre, i gruppi possono essere diversi, nella composizione, a seconda del metodo di clustering utilizzato.

Classificazione delle procedure/3

L'appartenenza delle entità ai gruppi potrebbe non essere esclusiva cioé che alcune unità possano ritrovarsi in più di un cluster (clumping, overlapping clusters).

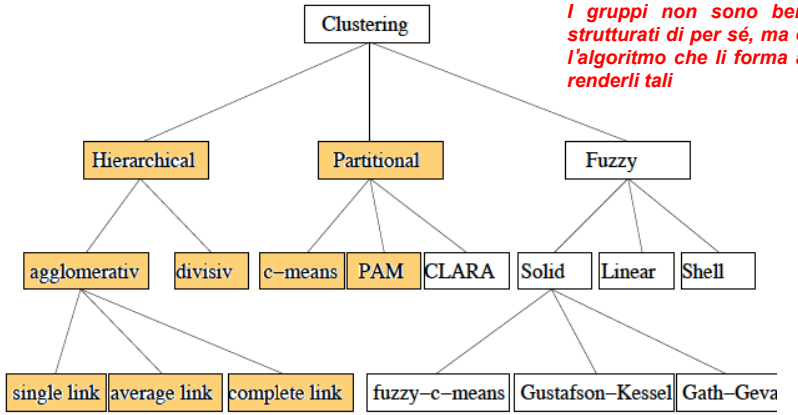
SFOCATA *(Fuzzy o soft)*

NETTA *(crispy o hard)*

Una situazione simile potrebbe riscontrarsi nella classificazione dei pazienti secondo i sintomi mostrati in quanto malattie diverse possono avere dei sintomi in comune.

Un altro esempio è la classificazione delle parole: uno stesso termine può avere significati diversi e deve essere collocato in gruppi diversi.

Classificazione degli algoritmi



I gruppi non sono ben strutturati di per sé, ma è l'algoritmo che li forma a renderli tali

Tuttavia, per la difficoltà di disporre di programmi di calcolo che gestiscano i cluster con sovrapposizione, adottiamo senza dubbio il modello di appartenenza esclusiva delle unità e discuteremo solo di tecniche in cui tutte le unità sono classificate e ciascuna è posta in uno ed un solo cluster.

Attenzione!

"The availability of computer packages of classification techniques has led to the waste of more valuable scientific time than any other "statistical" innovation (with the possible exception of multiple regression techniques)"
Cormack, 1970

I suspect that a major reason for the paper's popularity was the computer program. It was written in a highly portable dialect of FORTRAN, made pretty output pictures, was cheap to run, widely distributed (it was free!), and sinfully uncritical of its input data. Thus, many researchers could, and did, try the methods; those that liked the results could, and did, publish. Moreover, there were no burdensome significance tests to suggest that the data might not fit the clustering model. Thus, it was a fine way to obtain a computer's blessing without confronting the data's deficiencies.
S.C. Johnson (1985)

Criteri di scelta

Le caratteristiche di cui tenere conto nella scelta di un algoritmo di classificazione sono – indicativamente- le seguenti.

- **Matrice dei dati o matrice delle dissomiglianze**
- **Possibilità di applicazione a gradi data set**
- **Capacità di riconoscere cluster che hanno forma irregolare**
- **Gestione degli outliers**
- **Tempi di esecuzione**
- **Dipendenza dall'ordinamento delle unità nel data set**
- **Interpretabilità dei risultati**
- **Dipendenza da conoscenze a priori sul fenomeno che ha prodotto i dati**

Clustering gerarchica

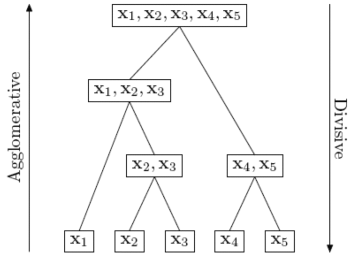
L'uso delle tecniche gerarchiche presuppone che sia possibile ricondurre lo studio di una matrice di dati ad una organizzazione gerarchica delle entità in gruppi annidati.

AGGREGATIVA

Fusioni a due a due di gruppi già formati. In fase iniziale ogni entità forma un gruppo e, come ultimo passo, tutte le unità sono in un gruppo unico

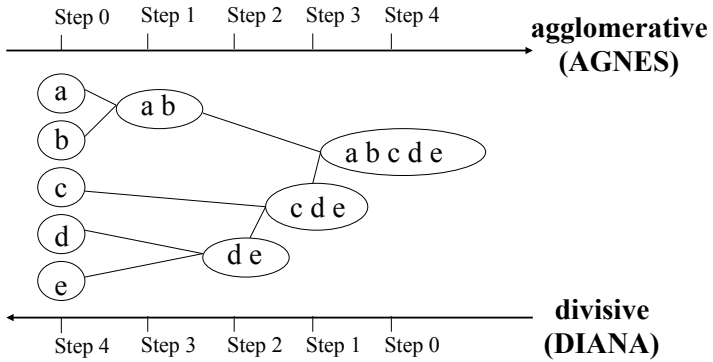
DIVISIVA

Divisioni a due a due di gruppi già formati. In fase iniziale tutte le entità sono in un unico gruppo e, come ultimo passo, le unità formano ciascuna un gruppo separato.



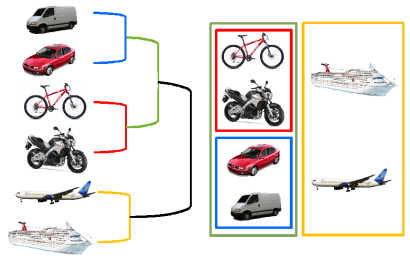
Clustering gerarchica/2

- **Si basa sulla matrice dei dati. Non è necessaro prefissare il numero dei gruppi, ma si deve predeterminare una condizione di stop nella formazione della gerarchia**



Clustering gerarchica aggregativa

Si fondono i due gruppi più prossimi in base alla prescelto indice di distanza o dissimilarità (detto legame) e si prosegue fino a formare un gruppo unico



Ad ogni passo della clustering gerarchica agglomerativa si fondono i due gruppi più prossimi secondo il legame che caratterizza l'algoritmo.

Il legame stabilisce come giudicare la dissomiglianza tra i gruppi ovvero distanza cluster-to-cluster

La maggiore debolezza di queste procedure è il vincolo di inclusione che costringe a proseguire con le aggregazioni già realizzate, senza poterle correggere, anche se ci si è accorti che non sono soddisfacenti.

Clustering gerarchica aggregativa/2

1. **Si parte dalla matrice delle distanze o dissimilarità.**
2. **Ad ogni passo si fondono i due gruppi più prossimi secondo il tipo di legame che caratterizza il metodo prescelto. Da notare che I singoletti sono cluster sebbene formati da un sola unità**
3. **Qui si forma una nuova matrice delle distanze o dissimilarità del nuovo gruppo con il resto dei gruppi (rimati invariati) usando SOLO i valori di quella ottenuta al livello precedente.**
4. **Si costruisce una nuova matrice con una riga ed una colonna in meno poché il gruppo appena formato è considerato come un singolo elemento**
5. **Si ritorna al punto 2 finché non si sia raggiunta la partizione ottimale ovvero tutte le unità sono confluite in un unico gruppo.**

Clustering gerarchica aggregativa/3

Il primo passo richiede O(n^2) operazioni per definire la matrice (nxn) delle distanze. L'ammontare delle elaborazioni dipende anche dal numero delle variabili e dalla loro scala di misurazione.

La matrice delle distanze-dissimilarità è considerata acquisita.

Nel secondo passo i calcoli dipendono dal tipo di legame. I tempi ora sono nell'ordine di: *O(n^2Log(n)*.

La gerarchia rihiede al massimo (n-1) passi o livelli di classificazione

$$C_j \neq \varnothing,\ 1 \le k \le n;\quad \bigcup_{i=1}^{k} C_j = D;\quad C_i \cap C_j = \varnothing \quad (i \neq j);$$

$$a)\ 1 \le n_i \le n-k+1,\ i = 1,2,\ldots,k \qquad b)\ \sum_{i=1}^{k} n_i = n$$

```
C   0.0000
E   2.1380 0.0000
J   2.2713 2.9579 0.0000
L   1.7135 2.3927 1.7772 0.0000
M   1.5460 1.9818 2.4575 1.0125 0.0000
N   2.6979 3.3566 1.9900 1.8520 2.6954 0.0000
O   2.9985 3.6848 3.4484 2.4272 2.6816 2.3108 0.0000
T   2.3859 2.3169 2.4666 1.4545 1.7581 2.2105 2.5041 0.0000
```

La clustering gerarchica ggregativa ha il difetto che per arrivare ai livelli superiori, di solito quelli più interessanti, occorre passare per diverse suddivisioni banali

Aggregazione dei gruppi

Si realizza valutatando le distanze Cluster-to-cluster cioè misure che riguardano due gruppi, anche se uno o entrambi sono dei singoletti cioè formati da una sola entità.

Ipotizziamo che due gruppi: "i" e "j" con n_i, e n_j enetità abbiano dissimilarità cluster-to-cluster $d_{i,j}$.

Supponiamo inoltre che $d_{i,j}$ sia la dissimilarità minore tra tutte le coppie di cluster cosicché il gruppo *i* ed il gruppo *j* sono quelli da fondere se si vuole formare un nuovo gruppo *k* con $n_k=(n_i+ n_j)$ elementi.

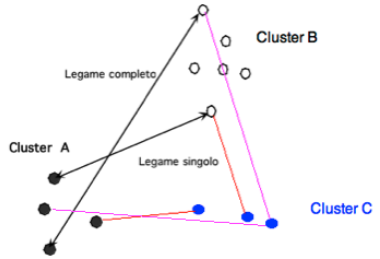
Se si opera con la matrice delle distanze, queste fornisce già la distanza cluster-to- cluster al primissimo livello.

Per fondere cluster con diversa cardinalità occorre definire il legame o link secondo cui esprimere la dissomiglianza tra due cluster.

Vari tipi di legami

LEGAME SINGOLO
La distanza o dissimilarità tra due cluster coincide con la distanza MINIMA tra due entità di cui una nel primo cluster e l'altro nel secondo.

LEGAME COMPLETO
La distanza o dissimilarità tra due cluster coincide con la distanza MASSIMA tra due entità di cui una nel primo cluster e l'altro nel secondo



Secondo il legame singolo si fondono A e C: secondo il completo si fondono A e B.

La distanza tra due specifiche entità fornisce la distanza cluster-to-cluster.

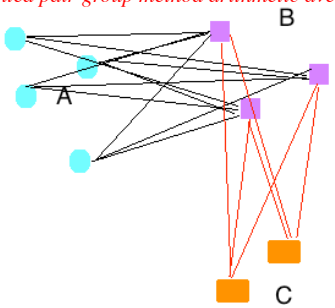
Vari tipi di legami/2

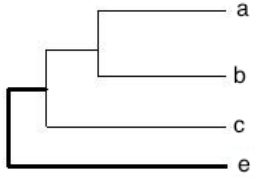
LEGAME MEDIO NON PESATO (UPGMA)
La distanza clustser-to-cluster è la distanza media tra tutte le unità contenute nei due cluster

LEGAME MEDIO PESATO (WPGMA)
La distanza cluster-to-cluster è la distanza media tra le unità dei due cluster tenuto conto di quante unità sono già contenute nei due cluster. Le unità inserite in cluster più in piccolo hanno un peso più piccolo.

(Un)Weighted pair group method average

Weighted pair group method arithmetic average





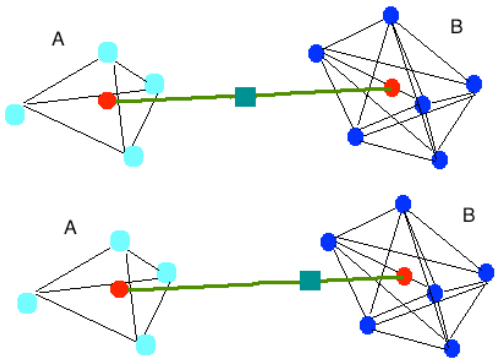
$$Unweighted: \quad d_{e\{a,b,c\}} = \frac{d_{a,e}+d_{b,e}+d_{c,e}}{3} = \frac{d_{a,e}}{3}+\frac{d_{b,e}}{3}+\frac{d_{c,e}}{3}$$

$$Weighted: \quad d_{e\{a,b,c\}} = \frac{\frac{d_{a,e}+d_{b,e}}{2}+d_{c,e}}{3} = \frac{d_{a,e}}{4}+\frac{d_{b,e}}{4}+\frac{d_{c,e}}{2}$$

Vari tipi di legami/3

DISTANZA TRA CENTROIDI (ignorano le dimensioni dei cluster da aggregare)

CENTROIDI PESATI (MEDIAN) (tengono delle numerodità dei cluster da aggregare



Un metodo non pesato tratta le unità in un cluster allo stesso modo

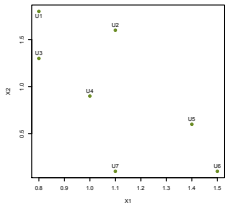
Un metodo pesato in realtà da lo stesso ad ogni cluster cosicché le entità nei cluster più piccoli, di fatto, pesano di più.

Il legame McQuitty (weighted average) è di questo tipo.

Legame singolo

Quadrati della distanza euclidea

	2	3	4	5	6	7
1	0.13	0.25	0.85	1.80	3.38	2.29
2		0.18	0.50	1.09	2.41	2.25
3			0.20	0.85	1.93	1.53
4				0.25	0.89	0.65
5					0.26	0.34
6						0.16



Le entità più prossime sono la U1 e la U2 che si fondono al livello di h=0.13 dove "h" è la soglia minimi di aggregazione formando il nuovo cluster (U1,U2)

	3	4	5	6	7
{1,2}	0.18	0.50	1.09	2.41	2.25
3		0.20	0.85	1.93	1.53
4			0.25	0.89	0.65
5				0.26	0.34
6					0.16

Il livello di aggregazione più prossimo è ora h=0.16 che coinvolge la U6 e la U7che risultano le più ravvicinate.

Legame singolo/2

$L_2 = [\{1,2\};\{3\};\{4\};\{5\};\{6,7\}]$

	3	4	5	{6,7}
{1,2}	0.18	0.50	1.09	2.25
3		0.20	0.85	1.53
4			0.25	0.65
5				0.26

Le dissimilarità si ottengono a livello dei singoli confronti entità/entità

	4	5	{6,7}
{1,2,3}	0.20	0.85	1.53
4		0.25	0.65
5			0.26

Se ad un dato livello più di una coppia di gruppi si candida per la fusione occorre procedere percorrendo tutte le possibilità.

Oppure si può anche estrarre a sorte

	5	{6,7}
{1,2,3,4}	0.25	0.65
5		0.26

$L_5 = [\{1,2,3,4,5\};\{6,7\}]$ per h=0.25 con d{1,2,3,4,5},{6,7}=0.26

$L_6 = [\{1,2,3,4,5,6,7\}]$.

Dendrogramma

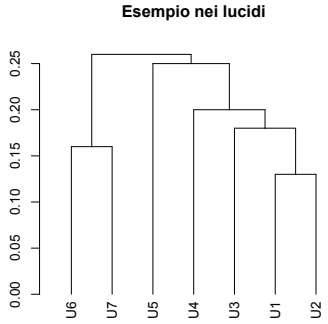
Costituisce la rappresentazioni grafica che più di frequente accompagna a esposizione di risultati relativi alla clustering gerarchica.

In orizzontale si riportano, equispaziate, le etichette delle unità soggette alla clustering

Sull'asse verticale si misura invece la distanza o dissimilarità.

Ad ogni aggregazione si forma un rettangolo aperto che ha base superiore al livello a cui avviene l'aggregazione e lati aperti congiunti con le unità aggregate poste sulle ascisse.

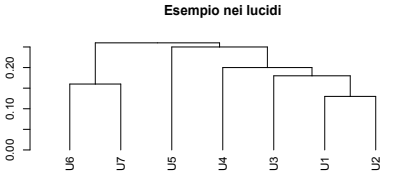
Ai livelli successivi lapertura dei rettangoli parte dal centro della base che unisce le entità di livello minore.

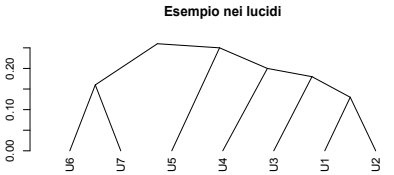


N.B. Una delle critiche più serie rivolte al dendrogramma è che esso non rivela una struttura, piuttosto ne impone una.

Forse non è il risultato di una elaborazione, ma è da considerare un sunto delle elaborazioni stesse

Dendrogramma/2





hclust di cluster (R) ordina I sotto-alberi in modo che le aggregazioni più recenti (unità più distanti) si collonano a sinitra.

La forma del dendrogramma può suggerire quanti gruppi esistono in effetti.

Le concatenazione descritte con i rettangoli sono più tradizionali.

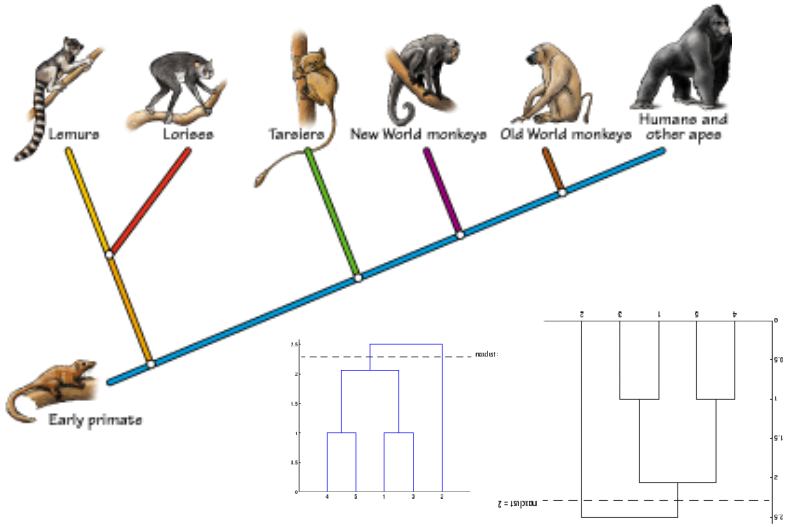
La descrizione con triangoli evoca maggiormente l'aspetto evolutivo dei gruppi

Da notare che il dendrogramma associato ad una gerarchia non è unico. Anzi, ne esistono $2^{(n-1)}$.

L'ordinamento ha importanza perché da esso dipende l'accostamento delle unità e quindi l'interpretazione dei risultati.

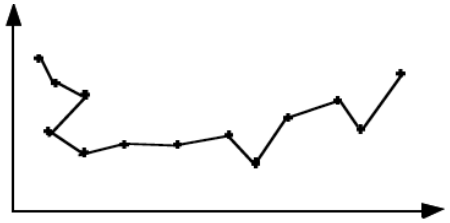
Dendrogramma/3

Lo schema del dendrogramma segue la stessa logica dell'evoluzione delle specie



Riflessioni sul legame singolo

Questa procedura basa la fusione tra due gruppi su di una sola coppia di entità e le altre vi entrano solo indirettamente. Il rischio è che, di minimo in minimo, si trovino nello stesso gruppo unità molto dissimili.



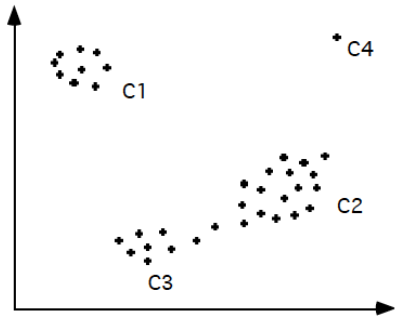
Il legame singolo tenta di formare una catena di entità, ognuna prossima allaltra, ma con una crescente dissimilarità tra quelle inserite subito e quelle inserite più tardi.

Il problema insorge dalla separatezza dei gruppi che può essere debole da impedire il buon funzionamento del legame singolo.

Riflessioni sul legame singolo/2

Tra gruppi ben separati possono trovarsi delle entità ibride che potrebbero agire da anello di congiunzione tra gruppi molto eterogenei.

Il legame singolo non ha difficoltà ad individuare il cluster del tipo C1 e C2, ma con clusters che hanno poca separazione nasce il chaining cioè la concatenazione per minimi di due cluster diversi.



E' un fenomeno sgradevole, ma frequente quando la struttura di gruppo non è molto pronunciata (cioè nei casi più interessanti).

Rispetto alla unità isolata (outlier) il legame singolo riesce ad evidenziarla molto bene

Legame completo (Johnson)

	2	3	4	5
1	7	1	9	8
2		6	3	5
3			8	7
4				4

	2	4	5
{1,3}	(7)	9	8
2		3	5
4			4

	{2,4}	5
{1,3}	9	8
{2,4}		5

	{2,4,5}
{1,3}	9

$d_{1,2} = 7,\ d_{3,2} = 6$

$\max\{6,7\} = 7 \Rightarrow d_{(1,3),2} = 7$

$L_0 = [\{u_1\};\{u_2\};\{u_3\};\{u_4\};\{u_5\};]$ *Tutti i cluster sono dei singoletti*

Al primo livello si fondono le unità u_1 ed u_3 dato che, a questo stadio, non esiste differenza tra legame singolo e legame completo.

Minima distanza massima

$h = 1$

$L_1 = [\{u_1, u_3\};\{u_2\};\{u_4\};\{u_5\};]$

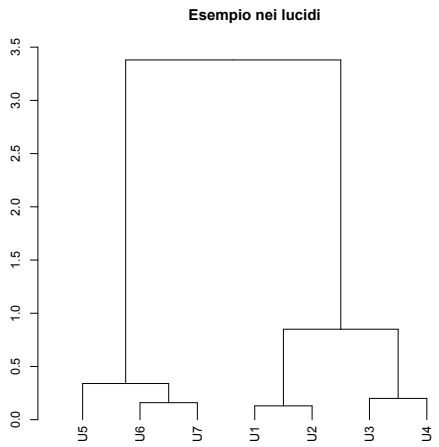
$h = 3, \quad L_2 = [\{u_1, u_3\};\{u_2 u_4\};\{u_5\}]$

h=3 è infatti la soglia più piccola che consente una fusione.

$L_4 = [\{u_1, u_3, u_2, u_4, u_5\}]$

Il livello h=5 consente la fusione di $\{u_5\}$ con $\{u_2,u_4\}$.Solo al livello h=9 i due clusters finali si fondono in un unico gruppo.

Legame completo/2



Il legame singolo ed il legame completo perseguono finalità diverse rispetto alle caratteristiche dei clusters: l'isolamento esterno e la coesione interna.

Il legame singolo trascura la coesione interna per preoccuparsi della separatezza dei gruppi.

Il legame completo cura maggiormente la compattezza di un cluster e non considera troppo l'isolamento rispetto agli altri clusters.

Legami medi (UPGMA e WPGMA)

$L_0 = [\{A\};\{B\};\{C\};\{D\};\{E\}]$

$h_1 = 1 \Rightarrow L_1 = [\{A,B\};\{C\};\{D\};\{E\}]$

il primo passo è sempre lo stesso; ciò che cambia è il ricalcolo della matrice delle distanze o dissimilarità. Ad esempio

$$d[\{A,B\};\{C\}] = (29+26)/2 = 27.5.$$

	B	C	D	E
A	1	29	50	50
B		26	49	53
C			5	13
D				4

Al livello successivo abbiamo

$h_2 = 4 \Rightarrow L_2 = [\{A,B\};\{C\};\{D,E\}]$

	C	D	E
{A,B}	27.5	49.5	51.5
C		5	13
D			4

La distanza

$$d[\{A,B\};\{D,E\}] = \frac{50+50+49+53}{4} = 50.5$$

$h_3 = 9 \Rightarrow L_3 = [\{A,B\};\{C\};\{D,E\}]$

	C	{D,E}
{A,B}	27.5	50.5
C		9

$h_4 = 42.83 \Rightarrow L_4 = [\{A,B,C,D,E\}]$

	{C,D,E}
{A,B}	42.83

Da notare che al livello L_q con soglia h_q abbiamo aggregato i gruppi C_i e C_j formando $C_i \cup C_j$. Per il livello successivo, L_{q+1}, la soglia h_{q+1} dovrà essere calcolata in base alla distanza tra $C_i \cup C_j$ ed ogni altro cluster, diciamo C_k, compreso nella partizione L_q.

Legami medi /2

La distanza tra C_k ed il cluster unione può essere schematizzata agevolmente:

$$d\left(C_k;\{C_i,C_j\}\right) = \frac{1}{n_k\left(n_i+n_j\right)}\sum_{\gamma_r=k}\sum_{\gamma_s=ioj} d\left(X_r,X_s\right)$$

$$= \frac{1}{n_k\left(n_i+n_j\right)}\sum_{\gamma_r=k} d\left(X_r,X_s\right) + \frac{1}{n_k\left(n_i+n_j\right)}\sum_{\gamma_r=k}\sum_{\gamma_s=j} d\left(X_r,X_s\right)$$

$$= \frac{n_i}{\left(n_i+n_j\right)} d(C_k,C_i) + \frac{n_j}{(n_i+n_j)} d(C_k,C_j)$$

Già disponibli

e quindi la distanza tra un cluster ed un cluster unione può essere calcolata in base alla matrice delle dissimilarità già ottenuta al livello L_q senza ricorrere alla matrice originale.

Legame medi/2

Ai fini del calcolo della distanza o dissimilarità tra il cluster unione $C_i \cup C_j$ e l'altro cluster C_k, il contributo dipende dalla numerosità dei cluster.

Pertanto, il cluster più grande tenderà a governare l'aggregazione ed i clusters più piccoli, specie se molto compatti, tenderanno a sparire.

Per questo motivo McQuitty (1960) ha proposto di migliorare la distanza cluster-to-custer dando lo stesso peso ai clusters che compongono il cluster unione.

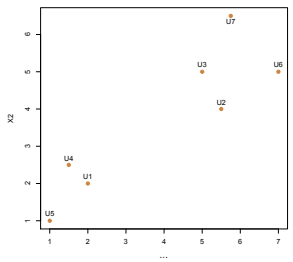
In particolare, il legame di McQuitty noto anche come WPGMA (metodo della media semplice), segue lo schema

$$d\left(C_k;\{C_i,C_j\}\right) = \frac{1}{2}d\left(C_k,C_i\right) + \frac{1}{2}d\left(C_k,C_j\right)$$

in cui la distanza tra il cluster esterno ed il cluster unione è pari ala media aritmetica semplice delle distanze tra cluster esterno e clusters componenti.

Esempio: McQuitty

distanza euclidea



entità	X_1	X_2
A	2.0	2.0
B	5.5	4.0
C	5.0	5.0
D	1.5	2.5
E	1.0	1.0
F	7.0	5.0
G	5.75	6.5

	B	*C*	*D*	*E*	*F*	*G*
A	4	4.2	0.7	1.4	5.8	5.9
B		1.1	4.3	5.4	1.8	2.5
C			4.3	5.7	2.0	1.7
D				1.6	6.0	5.8
E					7.2	7.3
F						2.0

$h_1 = 0.7$

$L_1 = [\{A,D\};\{B\};\{C\};\{E\};\{F\};\{G\}]$

	B	*C*	*E*	*F*	*G*
$\{A,D\}$	4.15	4.25	1.5	5.9	5.85
B		1.1	5.4	1.8	2.5
C			5.7	2.0	1.7
E				7.2	7.3
F					2.0

$h_2 = 1.1;$

$L_2 = [\{A,D\};\{B,C\};\{E\};\{F\};\{G\}]$

	$\{B,C\}$	*E*	*F*	*G*
$\{A,D\}$	3.3	1.5	5.9	5.85
$\{B,C\}$		5.55	1.9	2.1
E			7.2	7.3
F				2.0

Esempio continua

$h_3 = 1.5;$

$L_3 = [\{A,D,E\};\{B,C\};\{F\};\{G\}]$

	$\{B,C\}$	*F*	*G*
$\{A,D,E\}$	4.875	6.55	6.575
$\{B,C\}$		1.90	2.1
F			2.0

$h_4 = 1.90;$

$L_4 = [\{A,D,E\};\{B,C,F\};\{G\}]$

	$\{B,C,F\}$	*G*
$\{A,D,E\}$	5.712	6.575
$\{B,C,F\}$		2.05

$h_5 = 2.05;$

$L_5 = [\{A,D,E\};\{B,C,F,G\}]$

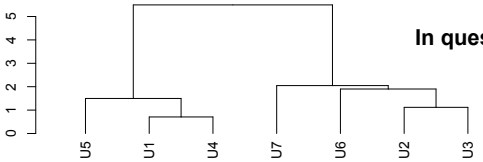
	$\{B,C,F,G\}$
$\{A,D,E\}$	6.143

Il legame McQyitty presume che i cluster abbiano le stesse dimensioni ovvero che la cardinalità dei gruppi non abbia rilevanza nella formazione dei nuovi gruppi.

Una conseguenza è che i gruppi meno numerosi tendono a pesare di più man mano che sono coinvolti nel processo di aggregazione.

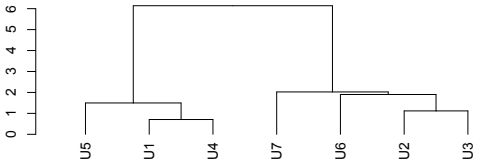
Esempio continua/2

Esempio nei lucidi/2. UPGMA



In questo caso il risultato è lo stesso

Esempio nei lucidi/2. WPGMA



Il metodo McQuitty è utile se si intende favorire la formazione di cluster della stessa dimensione.

Metodo di Ward

Il metodo di Ward (1963) considera tutte le possibili coppie di gruppi ad ogni stadio e procede alla fusione dei due che minimizzano l'incremento della devianza totale dal centroide del nuovo gruppo.

All'inizio del processo, quando ogni cluster è formato da un solo elemento, la devianza interna è zero

Quando due entità si fondono in un singolo cluster si introduce un grado di variabilità destinato a crescere in funzione della numerosità del gruppo stesso.

Wishart (1969) è riuscito ad esprimere il metodo di Ward come modifiche della matrice delle distanze.

$$d(i+j,k) = a_i d(i,k) + a_j d(j,k) + b d(i,j) + c\left|d(i,k) - d(j,k)\right|$$

Ward's Method (minimum variance)

$$a_i = \frac{n_k + n_i}{n_k + n_i + n_j} \quad a_j = \frac{n_k + n_j}{n_k + n_i + n_j} \quad b = -\frac{n_k}{n_k + n_i + n_j} \quad c = 0$$

Metodo di Ward/2

(1)	B	C	D	E	F	G
A	0.13	0.25	0.85	1.10	3.38	2.29
B		0.18	0.50	1.09	2.41	2.25
C			0.20	0.85	1.93	1.53
D				0.25	0.89	0.65
E					0.26	0.34
F						0.16

$$0.2433 = \left(\frac{2}{3}\right)0.18 + \left(\frac{2}{3}\right)0.25 - \left(\frac{1}{3}\right)0.13$$

$$0.8567 = \left(\frac{2}{3}\right)0.50 + \left(\frac{2}{3}\right)0.85 - \left(\frac{1}{3}\right)0.13$$

$h_1 = 0.13 \Rightarrow L_1 = \left[\{A,B\};\{C\};\{D\};\{E\};\{F\};\{G\}\right]$

Il livello h coincide con l'aumento della variabilità complessiva nei gruppi.

La valutazione del metodo di Ward avviene a posteriori invece che a priori come negli altri metodi.

(2)	C	D	E	F	G
{A,B}	0.2433	0.8567	1.8833	3.8167	3.4433
C		0.20	0.85	1.93	1.53
D			0.25	0.89	0.65
E				0.26	0.34
F					0.16

$h_2 = 016 \Rightarrow L_3 = \left[\{A,B\};\{C\};\{D\};\{E\};\{F,G\}\right]$

(3)	C	D	E	{F,G}
{A,B}	0.2433	0.8567	1.8833	5.3650
C		0.20	0.85	2.2533
D			0.25	0.9733
E				0.3467

$h_3 = 0.20 \Rightarrow L_3 = \left[\{A,B\};\{C,D\};\{E\};\{F,G\}\right]$

(4)	{C,D}	E	{F,G}
{A,B}	0.7250	1.8833	5.365
{C,D}		0.6667	2.32
E			0.3467

$h_4 = 0.3467 \Rightarrow L_4 = \left[\{A,B\};\{C,D\};\{E,F,G\}\right]$

	{E,F,G}
{A,B,C,D}	0.4975

$h_6 = 0.4975 \Rightarrow L_6 = \left[\{A,B,C,D,E,F,G\}\right]$

	{C,D}	{E,F,G}
{A,B}	0.7250	5.2833
{C,D}		2.1173

$h_5 = 0.7250 \Rightarrow L_5 = \left[\{A,B,C,D\};\{E,F,G\}\right]$

Metodo di Ward/3

L'impostazione del metodo di Ward si basa sulla coordinate cioè sulle modalità osservate delle variabili.

Come è noto, esiste un preciso legame tra coordinate e distanza euclidea al quadrato

$$\left(\tilde{x}_i - \tilde{x}_j\right)^t \left(\tilde{x}_i - \tilde{x}_j\right) = \tilde{x}_i{}^t \tilde{x}_i + \tilde{x}_j{}^t \tilde{x}_j - 2\tilde{x}_i{}^t \tilde{x}_j$$

$$= -2\left(-\frac{1}{2} d_{ij}^2\right) = d_{ij}^2$$

Il metodo di Ward può quindi essere applicato direttamente alla matrice di distanze/ dissomiglianze senza la mediazione della matrice dei dati.

Se la matrice delle distanze dissimilarità non è euclidea allora non possiamo stabilire una corrispondenza tra le distanze e le coordinate e quindi non dovremmo usare il metodo di Ward.

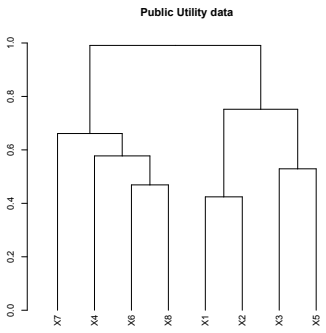
Esempio

Public utility data (H. E. Thompson) on 22 US public utility companies for the year 1975

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	1.00	0.64	−0.10	−0.08	−0.26	−0.15	0.04	−0.01
X_2	0.64	1.00	−0.35	−0.09	−0.26	−0.01	0.21	−0.33
X_3	−0.10	−0.35	1.00	0.10	0.44	0.03	0.11	0.01
X_4	−0.08	−0.09	0.10	1.00	0.03	−0.29	−0.16	0.49
X_5	−0.26	−0.26	0.44	0.03	1.00	0.18	−0.02	−0.01
X_6	−0.15	−0.01	0.03	−0.29	0.18	1.00	−0.37	−0.56
X_7	0.04	0.21	0.11	−0.16	−0.02	−0.37	1.00	−0.19
X_8	−0.01	−0.33	0.01	0.49	−0.01	−0.56	−0.19	1.00

	X1	X2	X3	X4	X5	X6	X7	X8
X1	0.00							
X2	0.42	0.00						
X3	0.67	0.57	0.00					
X4	0.68	0.67	0.67	0.00				
X5	0.61	0.61	0.53	0.70	0.00			
X6	0.65	0.70	0.70	0.60	0.64	0.00		
X7	0.69	0.63	0.67	0.65	0.70	0.56	0.00	
X8	0.70	0.58	0.70	0.50	0.70	0.47	0.64	0.00

$$d_{ij} = \sqrt{0.5\left(1 - r_{ij}\right)}$$



Il metodo di Ward tende ad aggregare cluster piccoli e tende a formare cluster della stessa numerosità.

La formula di Lance-Williams

I metodi aggregativi più comuni possono essere tutti ricondotti ad una unica formulazione.

l'aggiornamento della matrice delle distanze o dissimilarità (dopo che il cluster C_i ed il cluster C_j si sono fusi), da ogni altro cluster C_k può essere espresso con.

$$d_{h,k} = \alpha_i d_{h,i} + \alpha_j d_{h,j} + \beta d_{i,j} + \gamma \left| d_{h,i} - d_{i,j} \right|, \qquad d_{h,k} \geq 0$$

Dove i valori dei parametri $0 \leq \alpha_i, \alpha_j, \beta \leq 1,\ \gamma$ **determinano il legame.**

questi schemi sono detti COMBINATIVI perché riusano le distanze già calcolate rsultando perciò più rapidi di quelle che necessitano il ricalcolo.

Naturalmente non ci sono ragioni teoriche che impongono di proporre legami di aggregazione che ricadono in questa formulazione.

Se mai ragioni di opportunità, per sfruttare gli stessi schemi di calcolo e per facilitare i confronti tra schemi.

La formula di Lance-Williams/2

Name	(Reference)	α_i	β	γ
C1.	Single link (Florek *et al.*, 1951; Sneath, 1957)	$\frac{1}{2}$	0	$-\frac{1}{2}$
C2.	Complete link (McQuitty, 1960)	$\frac{1}{2}$	0	$\frac{1}{2}$
C3.	Group average link (Sokal and Michener, 1958; McQuitty, 1967)	$\frac{w_i}{(w_i+w_j)}$	0	0
C4.	Weighted average link (McQuitty, 1966, 1967)	$\frac{1}{2}$	0	0
C5.	Sum of squares (Jambu, 1978)	$\frac{(w_i+w_k)}{w_+}$	$\frac{(w_i+w_j)}{w_+}$	0
C6.	Incremental sum of squares (Ward, 1963; Anderson, 1966; Wishart, 1969)	$\frac{(w_i+w_k)}{w_+}$	$\frac{-w_k}{w_+}$	0
C7.	Centroid (Sokal and Michener, 1958; Gower, 1967b)	$\frac{w_i}{(w_i+w_j)}$	$\frac{-w_i w_j}{(w_i+w_j)^2}$	0
C8.	Median (Lance and Williams, 1966; Gower, 1967b)	$\frac{1}{2}$	$-\frac{1}{4}$	0
C9.	Flexible (Lance and Williams, 1966)	$\frac{1}{2}(1-\beta)$	$\beta(<1)$	0

N.B. In tabella si ha α_i=α_j

Legame flessibile

L'unico metodo flessibile menzionato in letteratura è quello con β=-0.25 che comporta la combinazione di coefficienti

$$\alpha_1 = \alpha_2 = \frac{1+0.25}{2} = 0.625, \quad \beta = -0.25, \quad \gamma = 0$$

	B	C	D	E
A	0.725	0.925	0.95	0.935
B		0.975	0.94	0.96
C			0.955	0.945
D				0.69

$h_1 = 0.69;$
$L_1 = [\{A\};\{B\};\{C\};\{D,E\}]$

	B	C	{D,E}
A	0.725	0.925	1.005625
B		0.975	1.015000
C			1.015000

$h_2 = 0.725;$
$L_2 = [\{A,B\};\{C\};\{D,E\}]$

	C	{D,E}
{A,B}	1.00625	1.081640625
C		1.015000

$h_3 = 1.00625;$
$L_3 = [\{A,B,C\};\{D,E\}]$

	D,E
{A,B,C}	1.058837891

$h_4 = 1.058837891; L_4 = [\{A,B,C,D,E\}]$

Per aggiornare bisogna tenere conto della formula:

$$d_{A(DE)} = \frac{5}{8}d_{AE} + \frac{5}{8}d_{AD} - \frac{2}{8}d_{DE}$$
$$d_{B(DE)} = \frac{5}{8}d_{BE} + \frac{5}{8}d_{BD} - \frac{2}{8}d_{DE}$$
$$d_{C(DE)} = \frac{5}{8}d_{CE} + \frac{5}{8}d_{CD} - \frac{2}{8}d_{DE}$$

$$d_{C(A,B)} = \frac{5}{8}d_{CA} + \frac{5}{8}d_{CB} - \frac{2}{8}d_{AB}$$
$$d_{(DE)(A,B)} = \frac{5}{8}d_{(DE)A} + \frac{5}{8}d_{(DE)B} - \frac{2}{8}d_{AB}$$
$$d_{(DE)(A,BC)} = \frac{5}{8}d_{(DE)(AB)} + \frac{5}{8}d_{(DE)C} - \frac{2}{8}d_{(AB}$$

Esempio comparativo

Matrice dei dati

$$\mathbf{x}_1 = (1.0 \quad 2.0 \quad 2.0)^T \qquad \mathbf{x}_4 = (3.0 \quad 4.0 \quad 3.0)^T$$
$$\mathbf{x}_2 = (2.0 \quad 1.0 \quad 2.0)^T \qquad \mathbf{x}_5 = (0.0 \quad 3.5 \quad 3.5)^T$$
$$\mathbf{x}_3 = (0.0 \quad 1.0 \quad 3.0)^T \qquad \mathbf{x}_6 = (2.0 \quad 2.5 \quad 2.5)^T$$

Quadrato della distanza euclidea

	1	2	3	4	5	6
1	0	2.0	3.0	9.0	5.5	1.5
2		0	5.0	11.0	12.5	2.5
3			0	18.0	6.5	6.5
4				0	9.5	3.5
5					0	6.0
6						0

Le parità tra le distanze in (3,5) e (3,6) non destano preoccupazione.

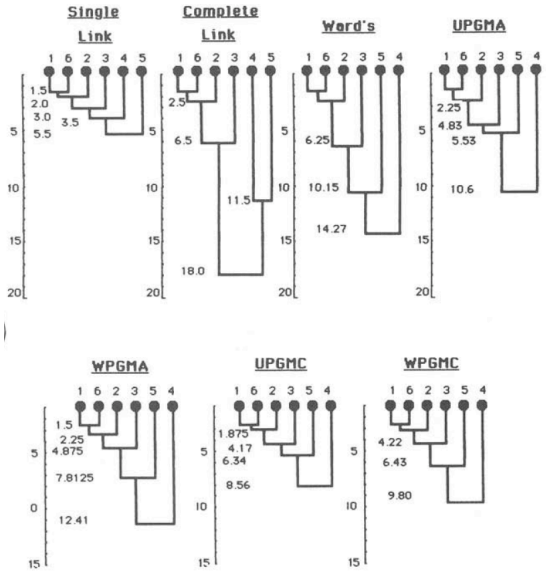
Ove possibile sarebbe opportuno migliorare la precisione delle misure per evitare le parità.

Esempio/continua

Alcuni studi comparativi tendono ad evidenziare una certa preferenza per il metodo di Ward.

Questo, almeno nella sua impostazione, presuppone una visione geometrica dei dati in uno spazio euclideo che non sempre ha riscontro nella realtà dei dati.

Inoltre ottimizza un criterio quadratico che potrebbe non essere quello più adatto per i dati in esame



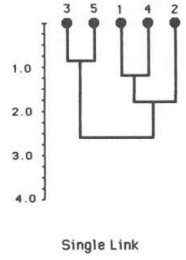
Esempio (finale)

	1	2	3	4	5
1	0	2.3	3.4	1.2	3.7
2		0	2.6	1.8	4.6
3			0	4.2	**0.7**
4				0	4.4
5					0

	1	2	3,5	4
1	0	2.3	3.4	**1.2**
2		0	2.6	1.8
3,5			0	4.2
4				0

	1,4	2	3,5
1,4	0	**1.8**	3.4
2		0	2.6
3,5			0

	1,2,4	3,5
1,2,4	0	**2.6**
3,5		0

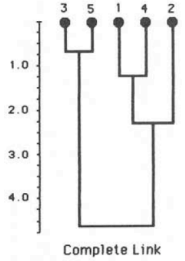


Single Link

	1	2	3,5	4
1	0	2.3	3.7	**1.2**
2		0	4.6	1.8
3,5			0	4.4
4				0

	1,4	2	3,5
1,4	0	**2.3**	4.4
2		0	4.6
3,5			0

	1,2,4	3,5
1,2,4	0	**4.6**
3,5		0

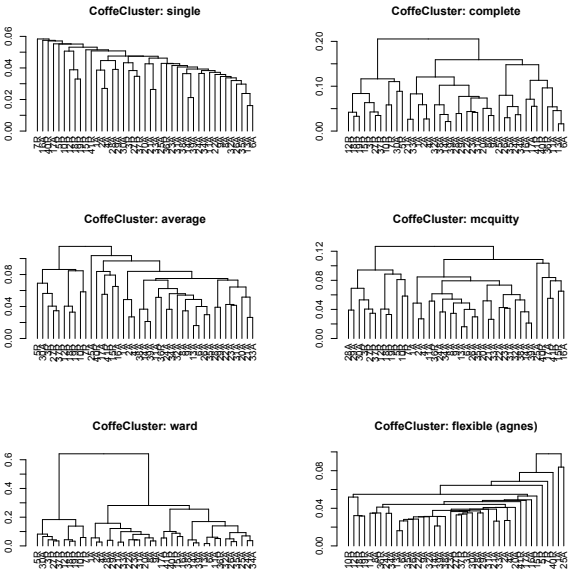


Complete Link

L'esito finale non è troppo diverso tra legame singolo e completo.

Non sempre si ritrova tale risultato e quando le entità sono numerose è necessario ricorrere a formule matematiche per interpretare i risultati della clustering

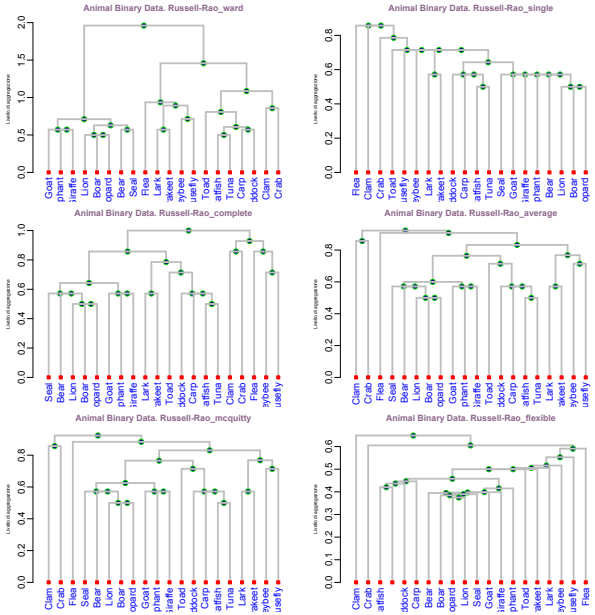
Applicazione: contenuto nutritivo dei cibi



Le soluzioni non sono equivalenti. Notate il concatenamento nel legame singolo.

Nel complesso sono riconoscibili quattro gruppi

Applicazione: variabili binarie

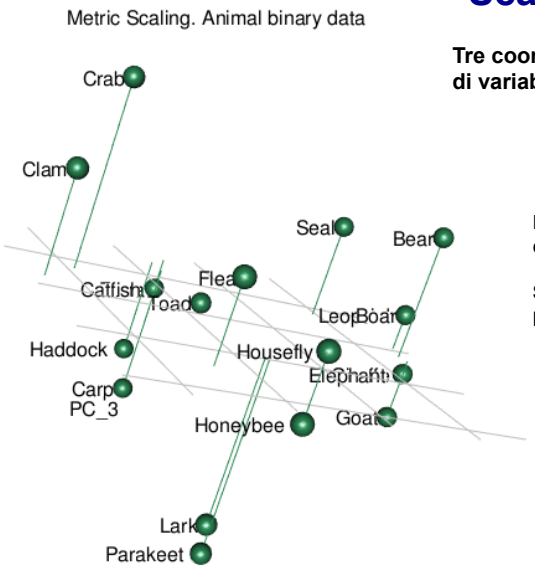


20 animali descritti da variabili binarie relative a loro caratteristiche

Le soluzioni sono diverse secondo il tipo di legame.

Nel complesso le 14 variabili binarie asimmetrice non sembrano sufficienti per formare gruppi credibili

Scaling metrico



Tre coordinate principali con il 46% di variabilità spiegata

L'effetto di parallasse rivela diversi microgruppi.

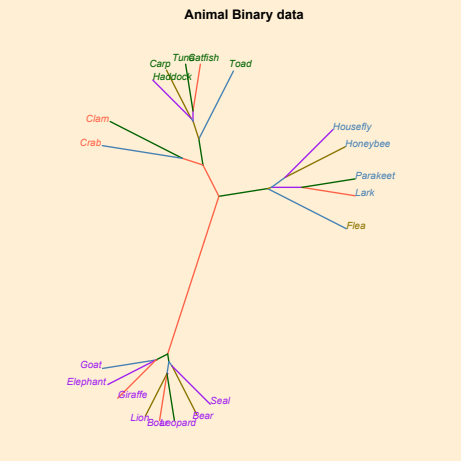
Sono necessarie altre variabili per una clustering più efficace

Alberi filogenetici

E' una rappresentazione alternativa della gerarchia che mostra i rapporti di somiglianza tra le unità

E' contenuto nel pacchetto ape in ambiente R e prevede diverse varianti.

Quella in esempio è tra le più seguite



Monotonicità

Ogni risultato dello schema Lance-Williams (1966-1967) dovrebbe produrreun dendrogramma che abbia almeno questo requisito minimo:

Ogni aggregazione avviene ad un livello di h superiore a quello usato per il livello precedente:

$$h_{q+1} > h_q \quad con \quad h_0 = 0.0$$

Per ottenere una sequenza monotona crescente è necessario che h_q= $\min\{d(C_i,C_j)\}$ con $C_i, C_j \in L_q$ sia monotono crescente con "q".

Supponiamo che al livello "q" si sia formato il cluster $C_i \cup C_j$ cioè

$$h_q = d(C_i, C_j) = \text{Min}\{d(C_r, C_s);\ C_r, C_s \in L_q\}$$

Al successivo livello di aggregazione abbiamo

$$h_{q+1} = \text{Min}\{d(C_r', C_s');\ C_r', C_s' \in L_{q+1}\} = d(C_i', C_j')$$

Se C_i' e C_j' sono allo stesso livello di quelli fusi al livello q la monotonicità è rispettata dato che al livello "q" furono fusi quelli con la distanza minima.

Monotonicità/2

Supponiamo che $C_r' = C_k$ e $C_j' = C_i \cup C_j$ siano uniti al livello "q", in base alla formula ricorsiva :

$$d_{k(i,j)} = \alpha_1 d_{ki} + \alpha_2 d_{kj} + \beta d_{ij} + \gamma \left| d_{ki} - d_{kj} \right| > d_{ij}$$

la monotonicità si ha se

$$1.\ \alpha_i + \alpha_j;\ 2.\ \alpha_i + \alpha_j + \beta \geq 1;\ 3.\ \gamma \geq -\min\{\alpha_i, \alpha_j\}$$

Quindi, il legame del centroide ed il legame mediano non sono monotoni dando luogo (sebbene di rado) inversioni nei livelli di aggregazione

Questo preclude la possibilità di ricostruire la matrice delle distanze a partire dai livelli di fusione (matrice cofenetica)

Pertanto sono ammissibili: singolo, completo, UPGMA, McQuitty, Ward, flessibile

Ultrametrica

Consideriamo la rappresentazione della gerarchia con il dendrogramma.

Ogni nodo o biforcazione ha un valore numerico che dovrebbe aumentare monotonicamente man mano che il nodo si allontana dalle prime aggregazioni.

Questo lo si può imporre sistematicamente richiedendo che i livelli di aggregazione formino una ultrametrica

$$d_{i,k} \leq \max\{d_{i,j}; d_{i,k}\}$$

Questa condizione equivale alle condizioni richieste ai parametri della formula di Lance-Williams per essere monotoni.

Scelta del legame

Alcune caratteristiche rendono problematico l'uso della classificazione gerarchica aggregativa

Le tecniche sono molto sensibili ai valori anomali ed ai disturbi nel calcolo delle distanze o dissimilarità. Ad esempio, come gestire le distanze uguali?

Non riescono ad intercettare bene strutture in cui sono compresenti gruppi poco numerosi e gruppi molto numerosi.

Anche forti divari nelle distribuzioni multivariate creano non poche difficoltà

Il legame singolo crea dei concatenamenti artificiali tra i cluster maggiori e favorisce la cannibalizzazione di quelli più piccoli.

Il legame completo forma cluster di tipo sferico anche se di questi non vi è traccia nei dati.

Il legame di Ward tende a formare cluster di eguale numerosità tendenzialmente piccoli.

Scelta del legame/2

Non è dato sapere quale sia la vera struttura dei dati.

Non è neanche possibile essere sicuri che una struttura ci sia in effetti.

L'esito finale della tecnica differisce secondo

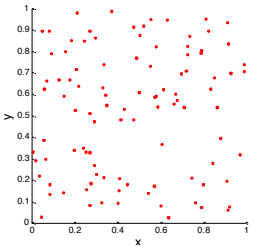
Da notare che una partizione in più gruppi verrà comunque ottenuta anche se il data set non si sogna minimamente di articolarsi in gruppi.

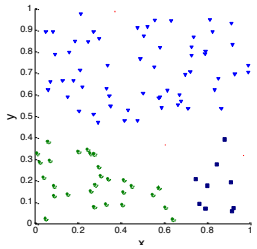
- **La dissimilarità/distanza usata**
- **La standardizzazione delle variabili**
- **La normalizzazione delle distanze**

L'analisi dei gruppi è dipende in modo stretto dall'applicazione in cui si applica ed è in buona misura effetto di scelte soggettive.

Cluster in dati casuali trovati da varie tecniche

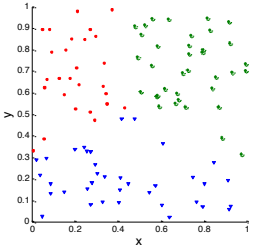
Random Points

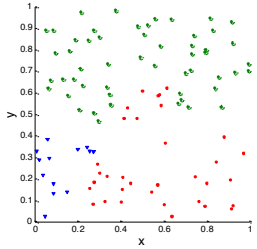




DBSCAN

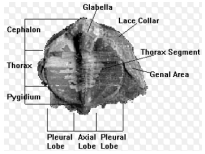
K-means





Complete Link

Matrice cofenetica

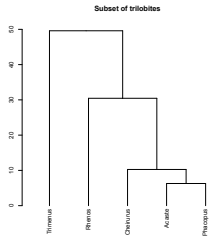


Phainò: apparire, rischiarare; Cum= insieme

Ogni procedura di clustering gerarchica aggregativa può essere considerata la trasformazione di una matrice di distanze in un dendrogramma o albero.

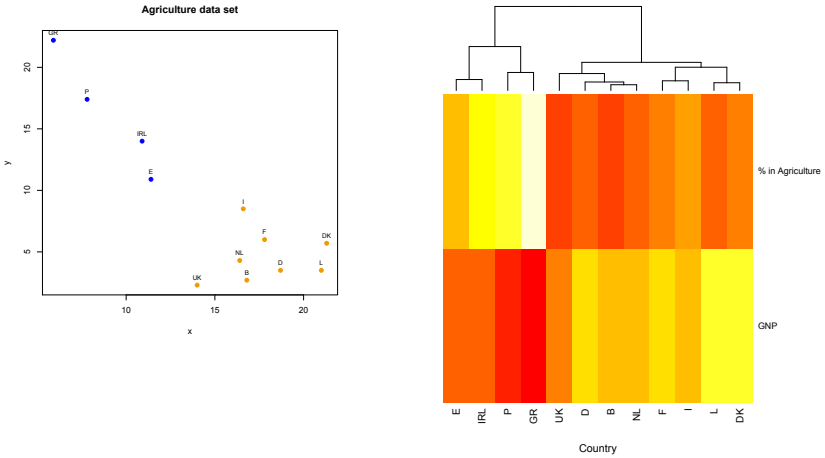
Genus	Body_Length	Glabella_Length	Gabella_Width
Acaste	23.14	3.5	3.77
Cheirurus	31.74	9.33	12.11
Phacopus	27.23	5.3	8.19
Rhenos	55.94	19	13.1
Trimenus	89.43	23.18	21.52

	Acaste	Cheirurus	Phacopus	Rhenos
Cheirurus	13.323			
Phacopus	6.285	7.207		
Rhenos	37.458	26.079	32.188	
Trimenus	71.391	60.071	66.077	34.784



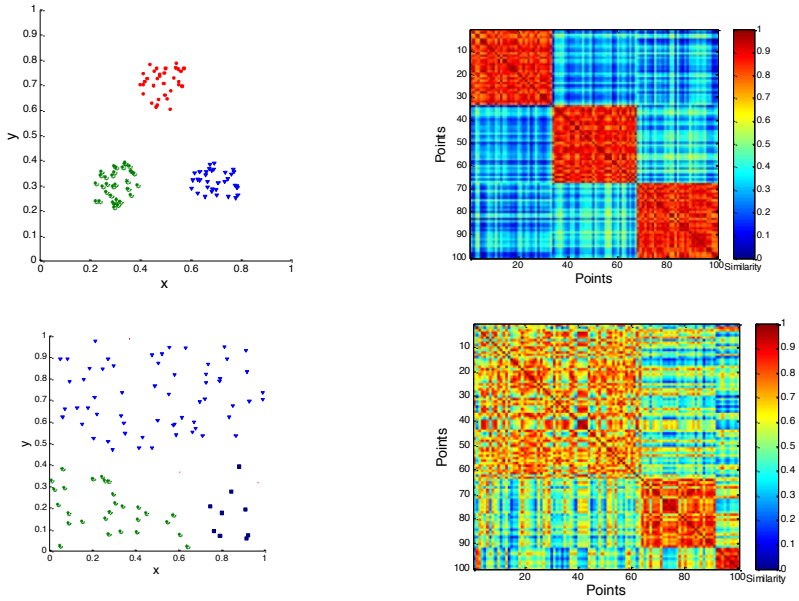
Il dendrogramma può formare a sua volta una matrice detta cofenetica che aiuta a valutare l'efficacia della classificazione.

Heat matrix (Esempio semplice)



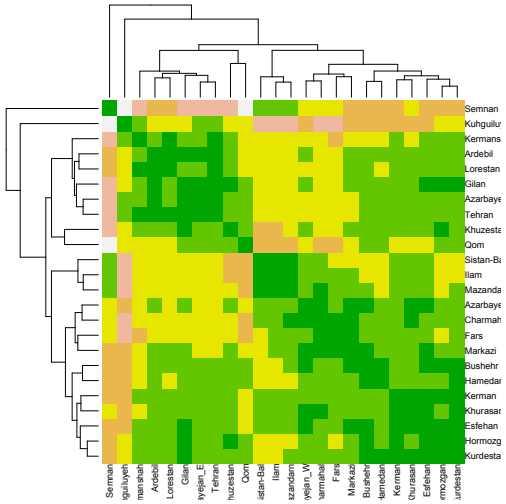
Si può notare la presenza di due gruppi di Paesi: 4+8

I cluster si trovano anche quando non ci sono



Heat matrix/2

Si cerca di rappresentare il dendrogramma in una matrice quadrata in cui l'intensità dei colori è un indicatore della prossimità tra le unità



Heat map of the distance matrix reordered according to the dendrogram.

Ricostruzione della matrice delle distanze

Siano U_i e U_j due unità; la distanza "ricostruita" tra di esse è il livello minimo h_q a cui U_i e U_j si trovano per la prima volta nello stesso cluster.

La riproduzione non è esatta. Se questo succede, si dice che il data set ha una struttura ad albero esatta e si presenta con cluster coesi al loro interno e ben separati rispetto all'esterno.

Poiché l'albero ha al massimo (n-1) livelli, la matrice cofenetica non potrà avere che (n-1) entrate distinte per cui alcune delle entrate avranno lo stesso valore.

	B	C	D	E	F	G	H
A	1	4	9	100	9	9	36
B		1	4	121	4	16	49
C			9	144	1	25	64
D				81	16	4	25
E					169	49	16
F						36	81
G							9

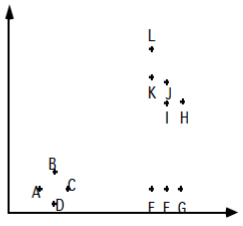
$h_1 = L_1\{A,B\}$
$h_2 = L_2\{C,F\};\{A,B\}$
$h_3 = L_3\{A,B,C,F\}$
$h_4 = L_4\{A,B,C,F\};\{D,G\}$
$h_5 = L_5\{A,B,C,F,D,G\}$
$h_6 = L_6\{A,B,C,F,D,G,H\}$
$h_7 = L_7\{A,B,C,D,E,F,G,H\}$

	B	C	D	E	F	G	H
A	1	1	4	16	1	4	9
B		1	4	16	4	4	9
C			4	16	1	4	9
D				16	4	4	9
E					16	16	16
F						4	9
G							9

Ecco la matrice ricostruita:
Per la ricostruzione possono bastare le relazioni d'ordine tra le distanze o dissimilarità.

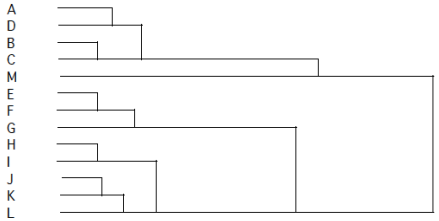
Matrice cofenetica/2

	B	C	D	E	F	G	H	I	J	K	L	M
A	1	2	2	7	8	8	12	11	12	12	12	10
B		1	1	6	7	7	11	10	11	10	11	7
C			2	5	6	6	10	9	10	9	10	8
D				5	6	8	12	11	12	11	12	8
E					1	3	7	6	7	6	7	13
F						2	6	5	6	7	8	14
G							4	5	6	7	8	14
H								1	2	3	4	10
I									1	2	3	9
J										1	2	3
K											1	7
L												6



La matrice cofenetica è equivalente all'albero di classificazione

Maggiore è la somiglianza tra la matrice cofenetica e la matrice delle distanze e meglio l'albero rappresenta il data set.



Esempio di matrice cofenetica

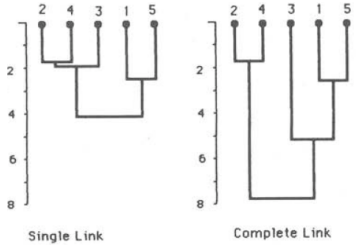
$$\mathcal{D}_2 = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{array} \begin{array}{c} \begin{array}{cccc} x_2 & x_3 & x_4 & x_5 \end{array} \\ \begin{bmatrix} 5.8 & 4.2 & 6.9 & 2.6 \\ & 6.7 & 1.7 & 7.2 \\ & & 1.9 & 5.6 \\ & & & 7.6 \end{bmatrix} \end{array}$$

Matrice delle distanze

$$\mathcal{D}_{Cs} = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{array} \begin{array}{c} \begin{array}{cccc} x_2 & x_3 & x_4 & x_5 \end{array} \\ \begin{bmatrix} 4.2 & 4.2 & 4.2 & 2.6 \\ & 1.9 & 1.7 & 4.2 \\ & & 1.9 & 4.2 \\ & & & 4.2 \end{bmatrix} \end{array}$$

$$\mathcal{D}_{Cc} = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{array} \begin{array}{c} \begin{array}{cccc} x_2 & x_3 & x_4 & x_5 \end{array} \\ \begin{bmatrix} 7.6 & 5.6 & 7.6 & 2.6 \\ & 7.6 & 1.7 & 7.6 \\ & & 7.6 & 5.6 \\ & & & 7.6 \end{bmatrix} \end{array}$$

Sia il legame singolo che completo applicati alla matrice cofenetica riproducono lo stesso albero ottenibile con la matrice originale delle distanze.



Correlazione cofenetica

E' possibile quantificare la qualità di una struttura ad albero sintetizzando il confronto tra la matrice originale delle distanze e la matrice delle distanze ricotruite ottenuta dal dendrogramma (matrice cofenetica).

Tale confronto coinvolge n(n-1)/2 valori d_{ij} e d'_{ij}. La misura più nota è il coefficiente di correlazione tra le osservate d_{ij} e quelle ricostruite d'_{ij}

La sintesi del confronto si può ottenere con il coefficiente di correlazione cofenetico (il classico indice di Pearson) tra le distanze nelle corrispondenti posizioni delle due matrici (diagonale esclusa) usando un solo lato della matrice.

$$r(d,d') = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(d_{ij}-\bar{d}\right)\left(d_{ij}'-\bar{d}'\right)}{\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(d_{ij}-\bar{d}\right)^2}\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(d_{ij}'-\bar{d}'\right)^2}}; \quad \bar{d} = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_{ij}}{n(n-1)}; \quad \bar{d}' = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_{ij}'}{n(n-1)}$$

Correlazione cofenetica/2

Lo si può ritenere un indizio di come la struttura gerarchica dei dati, se presente, potrebbe essere riprodotta dal dendrogramma dal quale sono state ricostruite le distanze.

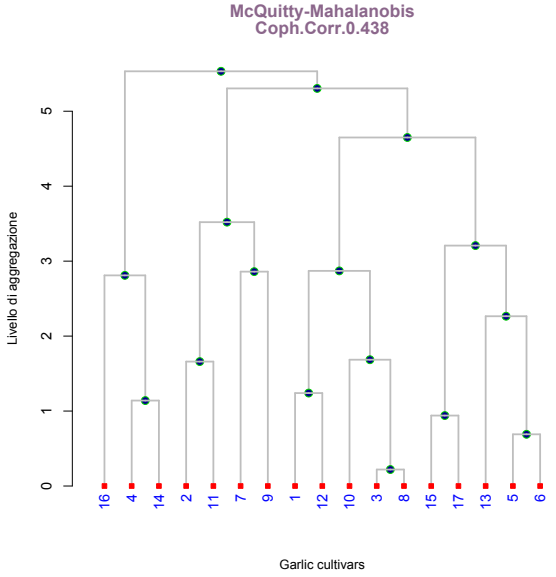
La correlazione cofenetica è tendenzialmente positiva (perché riferita ad ordinamenti simili) ed in genere elevata anche per ricostruzioni molto approssimative.

Solo i valori molto alti, diciamo superiori a 0.90 (meglio 0.95) possono essere considerati realmente significativi.

In generale, r(d,d') è pari ad uno quando ma matrice delle distanze incorpora dei gruppi ben strutturati.

Valori ridotti di r(d,d') , diciamo inferiori a 0.70, rendono contestabile la riproduzione della matrice delle distanze o dissimilarità e, forse, il data set non ha clusters o non ha gruppi nella forma che il legame adoperato è in grado di riconoscere.

Esempio



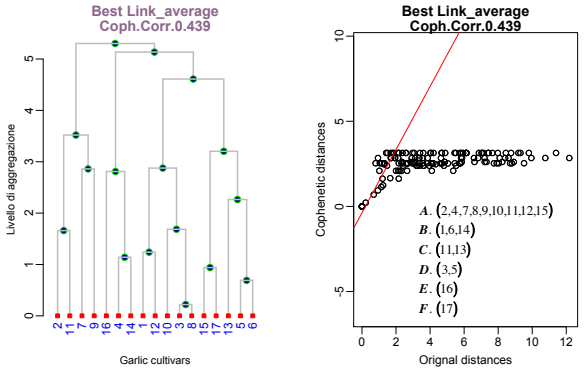
L'input è una matrice di distanze di Mahalanobis tra 17 cultivar di aglio (Da Silva, 2013)

La discrezione è solo sul legame. Qui abbiamo scelto il McQuitty.

I gruppi trovati dagli autori sono non hanno molto riscontro dal grafico ed infatti la correlazione cofenetica è basa

A. (2,4,7,8,9,10,11,12,15)
B. (1,6,14)
C. (11,13)
D. (3,5)
E. (16)
F. (17)

La migliore offerta

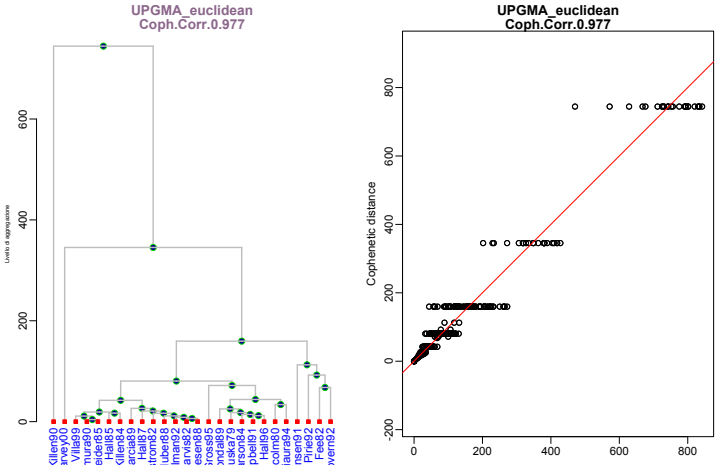


Il diagramma di Shepard è un diagramma di dispersone (o scatterplot) dove sulle ascisse sono riportate le distanze originali e sulle ordinate quelle approssimate. In questo caso quelle ricostruite in base al legame clustr-to-cluster.

Il legame più efficace (nei termini della correlazione cofenetica) è quello delle medie non ponderate. Forse la clustering gerarchica agglomerativa non è adatta per questi dati.

La migliore offerta/2

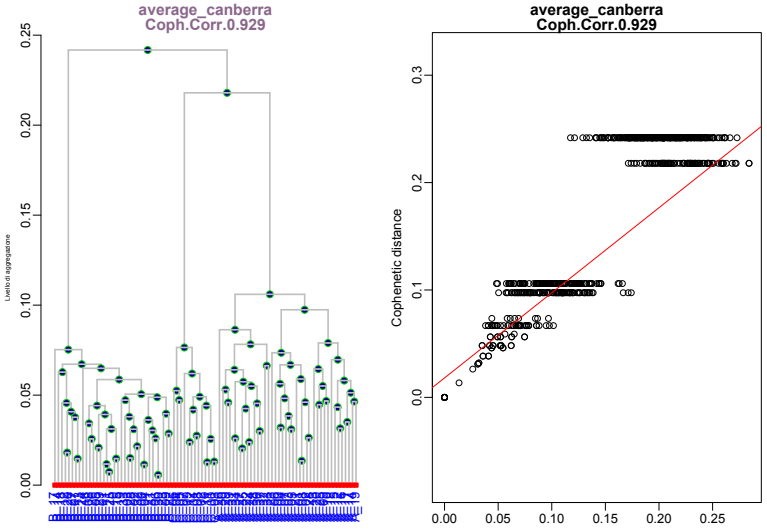
Qualora fosse disponibile la matrice dei dati, l'escussione può riguardare anche la tipologia di distanza.



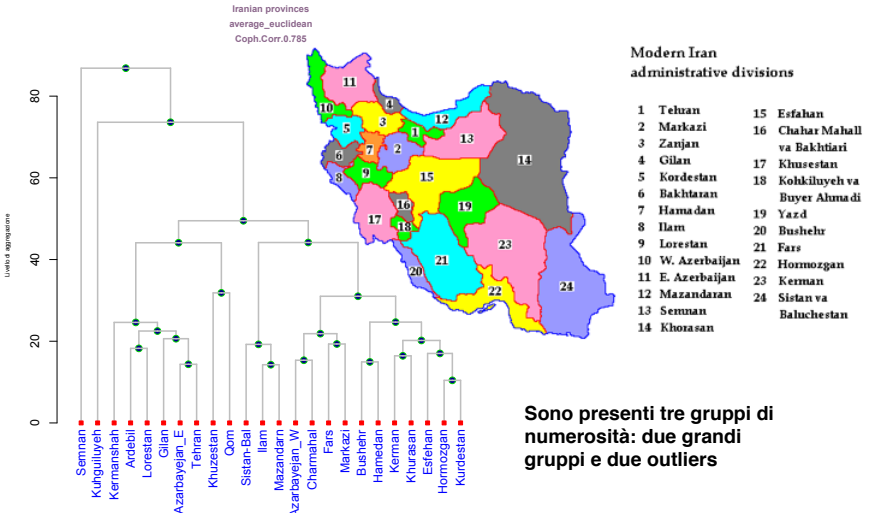
Combinando I comandi vegdist, hclust(+agnes) abbiamo applicato 48 metodi al data set smoking del package Hsau.

Applicazione: Chernoff data set

Sono presenti tre gruppi di numerosità: 42, 12, 22 dobbiamo validare la soluzione



Applicazione: Iranian provinces



Sono presenti tre gruppi di numerosità: due grandi gruppi e due outliers

Valutazione preliminare (ground truth)

Per sottoporre a verifica una tecnica o una sua variante si adopera una data set di cui si conosce da altre fonti la situazione generale

In molti contesti applicativi si adopera il termine *ground truth* (verità al suolo) per indicare la valutazione di una procedura su di una parte del data set al fine di metterne a punto i parametri le opzioni.

La classificazione delle unità del cosiddetto data set di addestramento (training data set) richiede la loro disamina accurata e l'accertamento univoco della categoria di appartenenza.

Nel remote sensing il satellite valula la luce riflessa da vegetazione, acqua, edifici, ma la "firma spettrale" delle varie entità deve essere identificata su zone conosciute

Matrice di confondimento

Il confronto tra la categoria a cui appartengono le unità e quella loro attribuita dalla procedura si può sintetizzare nella matrice di confondimento

Clustering stimata \ Clustering teorica	1	0	
1	a	b	$a+b$
0	c	d	$c+d$
	$a+c$	$b+d$	$a+b+c+d$

a) La unità i e la j sono nello stesso cluster sia nella stimata che nella teorica

b) La i e la j sono nello stesso cluster nella stimata ma non nella teorica

c) La i e la j sono nello stesso cluster nella teorica ma non nella stimata

d) La i e la j non sono mai nello stesso cluster

La matrice di confondimento ottenuta da una procedura di classificazione può dare una idea della qualità della clustering usando uno dei tanti indici di associazione binaria.

Preferenza per quelli che coinvolgono "d" in quanto tanto l'appartenenza che la non appartenenza contribuiscono a definire la natura dei cluster.

Indice di Rogers-Tanimoto

$$Q = \frac{a+d}{a+d+2(b+c)}$$

L'indice è pari ad uno se tutte le coppie ricadono nello stesso cluster sia del raggruppamento stimato che in quello teorico.

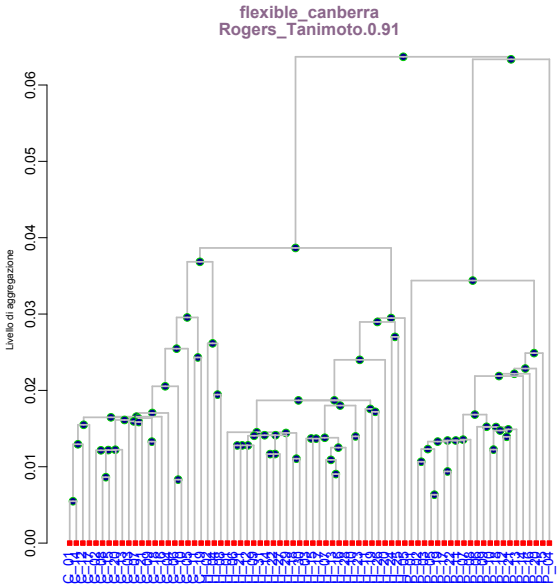
L'indice diminuisce e tende a zero man mano che le coppie sono allocate in cluster diversi nei due raggruppamenti.

Naturalmente è possibile che gli accoppiamenti nello stesso cluster siano sbagliati in entrambi i raggruppamenti.

Non si ammette la possibilità di singoletti.

Niente si dice ad esempio sulle terne di unità che potrebbero essere disarticolate nelle due classificazioni a confronto.

Esempio: Lubishew data set 1

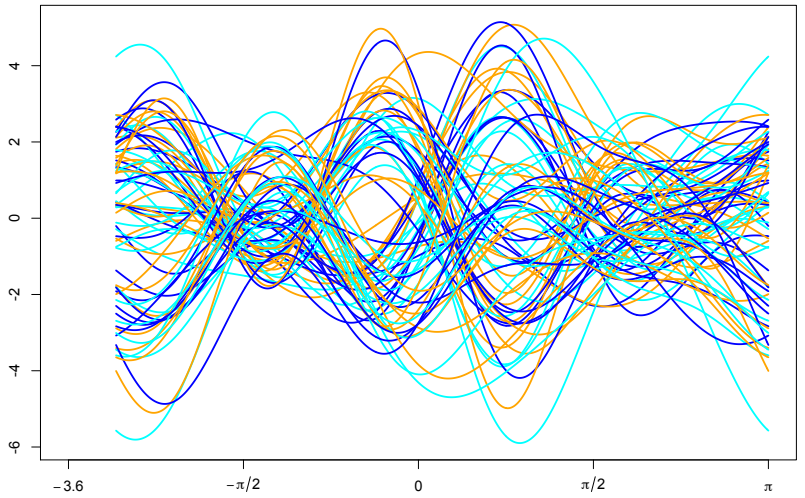


MEASUREMENTS ON MALES
Chaetocnema concinna

	1	0	
1	781	781	906
0	0	1795	1795
	781	1920	2701

I tre gruppi sembrano ben individuati. Si può generalizzare?

Curve di Andrews



I gruppi sono distinguibili: tre di cui due molto simili.

Indice di Rand (Rand index)

William M. Rand, nel 1971, ha proposto una statistica che quantifica la tendenza al raggruppamento di un certo data set, quale emerge come esito di una procedura di clustering.

Rand è partito dalle seguenti considerazioni:

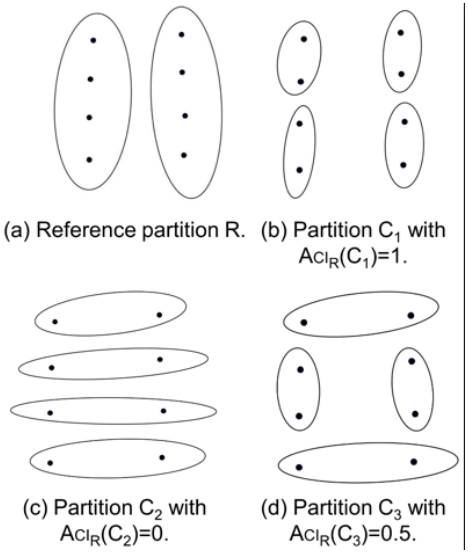
- **Ogni entità ricade in uno ed un solo cluster.**
- **Il carattere del cluster è definito sia dalle entità che contiene che da quelle che non contiene.**
- **Tutte le entità hanno la stessa importanza nel definire la qualità di una partizione.**

Indice di Rand /2

Rand dedusse che l'unità di confronto fondamentale tra due clusters fosse costituita dalla collocazione delle entità considerate due a due:

Se una coppia di entità è inclusa in uno stesso cluster da duedistinte classificazioni questo è un segno di similarità tra le due ed è un segno di dissimilarità il fatto che vengano abbinate in una classificazione e poi spaiate in un'altra.

Rand non considera le entità che risultino scompagnate in entrambe le classificazioni.



(a) Reference partition R. (b) Partition C_1 with $Acl_R(C_1)=1$.

(c) Partition C_2 with $Acl_R(C_2)=0$. (d) Partition C_3 with $Acl_R(C_3)=0.5$.

Rand Index/3

Il test di Rand è dato dalla formula

$$Rand = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_{ij}}{\binom{n}{2}} \quad con \quad c_{ij} = \begin{cases} 1 \ se\ U_i\ e\ U_j\ sono\ nello\ stesso\ cluster \\ \quad sia\ nella\ teorica\ che\ nella\ stimata \\ 1 \ se\ U_i\ e\ U_j\ sono\ in\ cluster\ diversi \\ \quad sia\ nella\ teorica\ che\ nella\ stimata \\ 0\ otherwise \end{cases}$$

In pratica si tratta del coefficiente di associazione per binarie simmetriche detto *simple matching*.

$$SM = \frac{a+d}{a+b+c+d}$$

Il test di Rand varia tra zero: quando una classificazione coincide con la partizione disgiunta e l'altra con quella congiunta; ed uno: quando le due classificazionicoincidono perfettamente.

Adjusted Rank Index

Una evidente difficoltà di usare l'indice di Rand è che il suo valore atteso nel confronto di due classificazioni formate a caso non è sempre lo stesso (diciamo pari a zero) come dovrebbe.

La versione del Rand index proposta da Hubert and Arabie (1985) ipotizza un modello di casualità basato sulla distribuzione ipergeometrica

Cioè le clustering casuali condividono comunque un numero di cluster fisso e altrettanto fissato è pure il numero di unità che contengono.

Partiamo dalla tabella di contingenza relative a due classificazione (teorica e stimata)

Teorica/Stimata	C_1	C_2	$\cdots$	C_k	
C_1	n_{11}	n_{12}	$\cdots$	n_{1k}	$n_{1.}$
C_2	n_{21}	n_{11}	$\cdots$		$n.$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
C_k	n_{k1}	n_{11}	$\cdots$	n_{11}	$n_{k.}$
	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.k}$	n

Adjusted Rank Index/2

Uno schema generale di un indice che abbia un valore atteso costante è

$$\frac{Indice - Valore\ \ atteso}{Valore\ \ massimo - Valore\ \ atteso}$$

Che ha limite superiore pari ad uno ed ha valore atteso nullo quando le due classificazioni sono da considerarsi formate a caso.

Si può dimostrare che l'indice di Rand <u>Adjusted</u> è pari a:

$$\frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\right] - \left[\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}\right] / \binom{n}{2}}$$

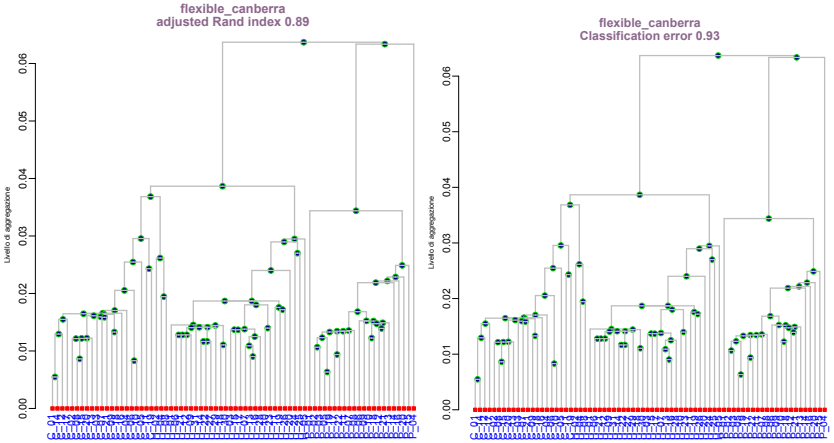
Class \ Cluster	v_1	v_2	v_3	*Sums*
u_1	1	1	0	2
u_2	1	2	1	4
u_3	0	0	4	4
Sums	2	3	5	$n = 10$

$$a = 7;\ b = \binom{2}{2} + \binom{4}{2} + \binom{7}{2} - 7 = 6$$

$$c = \binom{2}{2} + \binom{3}{2} + \binom{5}{2} - 7 = 7;\ d = \binom{10}{2} - 7 - 6 - 7 = 25$$

$$Ri = \frac{7+25}{45} = 0.711,\ aRi = \frac{7 - 14*13/45}{(14+13)/2 - 14*13/45} = 0.313$$

Esempio



Tasso di errore

In alternativa all'ARI si può usare il tasso di errore (complemento ad uno dell'indice di Russell-Rao)

$$ClassErr = 1 - \frac{a}{a+d+b+c}$$

```
> a
[1] 1 2 3 1 2 3 1 1 1
> b
[1] "A" "B" "C" "A" "B" "C" "A" "B" "C"
> classError(a, b)
$misclassified
[1] 8 9

$errorRate
[1] 0.2222222
```

Numero di cluster

La scelta del numero dei cluster k deve essere ben ponderata. In genere è compreso tra un limite minimo di 3-5 (per evitare cluster troppo ampi) ed un massimo (cluster troppo piccoli vanificherebbero il processo di semplificazione che è alla base del raggruppamento).

In generale, più piccoli sono i cluster più informativa risulta la collocazione di una unità rispetto alla tipologia descritta dal clsuter

D'altra parte, classi numerose possono dar luogo a profili irregolari complicando l'uso della cluster nelle applicazioni.

Il numero delle classi è perciò un compromesso tra esigenze contrastanti: da un lato c'è la necessità di trascurare i dettagli non necessari, dall'altro si vuole evitare la perdita di informazioni preziose.

Occorreranno diverse prove prima di pervenire ad una scelta soddisfacente.

Numero di cluster/2

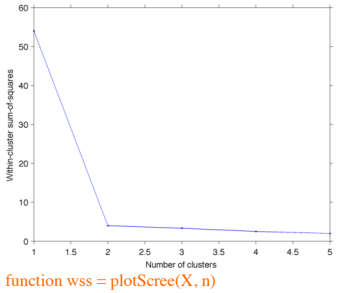
L'impiego delle tecniche gerarchiche è già di per sé una dichiarazione di incertezza sul numero dei gruppi.

Il problema è di capire quando fermare lo sviluppo dell'albero.

Se si intravvede uno spazio molto grande tra un livello ed un altro è lì che bisogna cercare il numero di gruppi ottimale in quanto qui forse si mischiano gruppi molto diversi.

In questo senso può essere d'aiuto un grafico che abbia sulle ascisse il numero di gruppi k e sulle ordinate il livello di aggregazione necessario per formare il k-esimo gruppo.

Un marcato appiattimento della curva, individuato con ampie riserve di soggettività, indicherà il valore giusto di k.




```
function wss = plotScree(X, n)
wss = zeros(1, n);
wss(1) = (size(X, 1)-1) * sum(var(X, [], 1));
for i=2:n
    T = clusterdata(X,'maxclust',i);
    wss(i) = sum((grpstats(T, T, 'numel')-1) .*
sum(grpstats(X, T, 'var'), 2));
end
hold on
plot(wss)
plot(wss, '.')
xlabel('Number of clusters')
ylabel('Within-cluster sum-of-squares')
```

Average silhuette width

Come misura della coesione e separatezza di un cluster Kaufman and Rousseeuw (1990) hanno proposto il coefficiente

$$s_{i,r} = \frac{b_{i,r} - a_{i,r}}{\max\left(b_{i,r}, a_{i,r}\right)},\ i = 1,2,\cdots,n;\ r = 2,3,\cdots,k$$

$$s.c.\quad a_{i,r} = \frac{\sum_{s=1}^{n} d_{i,s}}{\sum_{s=1}^{n} I(\gamma_s = \gamma_i)}\ ;\ d.c.\ b_{i,s} = \frac{\sum_{r=1}^{n} d_{i,s}}{\sum_{s=1}^{n} I(\gamma_s \neq \gamma_i)}$$

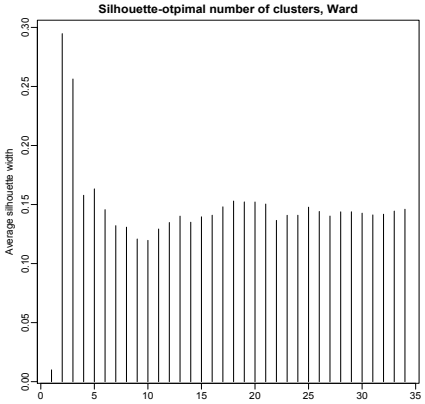
Se s_i è prossimo ad uno allora l'unità i-esima è inserito correttamente nel suo cluster. Se invece la silhuette tende a zero si deve sospettare che l'unità sia stata collocata nel cluster sbagliato.

La media delle s_i fornisce un criterio per la scelta di k

Il numero dei cluster presunto è quello che corriponde al massimo valore della media delle silhuette.

I valori elevati evidenziano una clusterring ben strutturata.

Esempio: sparrow data



Succede spesso che questa strategia suggerisca k=2 come numero ottimale di gruppi.

Se si esclude questa possibilità forse dovuta alla debolezza del metodo si può optare per k=3

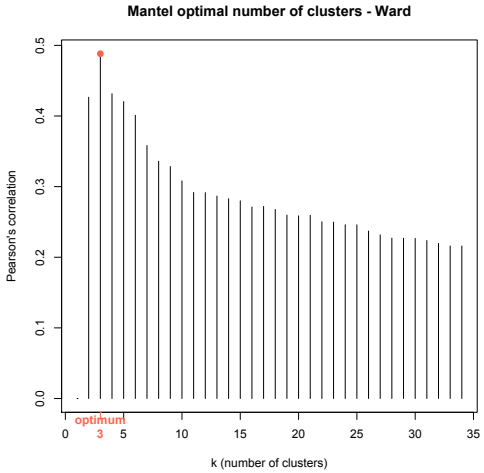
Nel caso di questi dati il numero più indicato è k=4 che non è facilmente individuabile nella silhuette.

Esempio: sparrow data/2

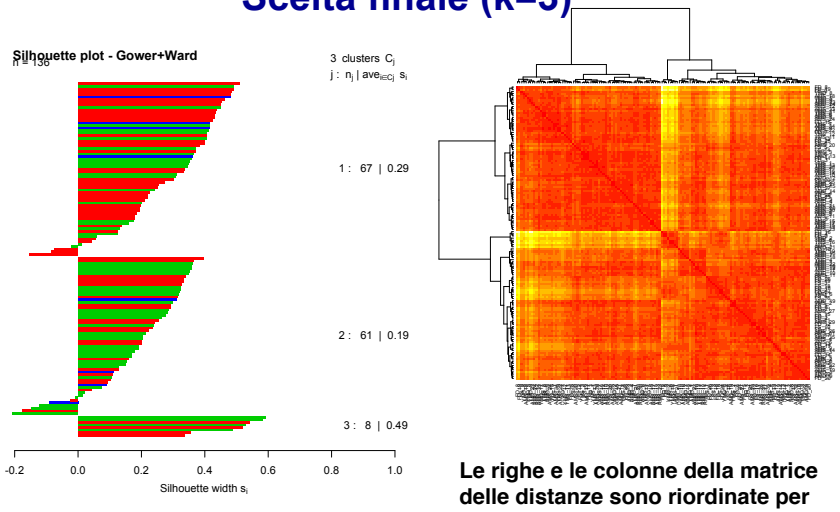
Un'altra possibilità è di utilizzare il coefficiente di correlazione cofenetica tra matrice originale delle distanze e matrice ricostruita in base alla classificazione in k gruppi

In questo caso c'è una indicazione per k=3 che sembra confermare ls soluzione con la average silhuette width.

k=4 non è un'opzione con questa strategia



Scelta finale (k=3)



Le righe e le colonne della matrice delle distanze sono riordinate per evidenziare l raggruppamenti più evidenti. I valori sono sostituiti a colori graduati secondo l'intensità della distanza.

Indice di Dunn

E' un criterio di validità della clustering proposto da Dunn (1973)

E' definito in base al rapporto tra la minima separatezzza tra due cluster in rapporto al massimo del diametro dei due cluster

$$\text{Dunn's index} \quad \min_{1\le r\le k}\left\{\frac{\min\limits_{1\le r\le k,\,1\le s\le k}\left\{\Psi(C_r,C_s)\right\}}{\max\limits_{1\le r\le k,\,1\le s\le k}\left\{\Delta(C_r,C_s)\right\}}\right\}$$

Esistono diverse alternative per questo indice legate al modo di scegliere la separatezza tra i cluster c("single", "complete", "average", "hausdorff") **e la separatezza tra due cluster** ("complete","average")

Valori elevati dell'indice di Dunn sono un indizio dell'esistenza di gruppi compatti e ben separati.

Il numero di cluster k è indicato nel punto di massimo dell'indice

Indice di Davies-Bouldin

E' stato proposto da Davies and Bouldin (1979) deriva da una formula di compromesso tra compattezza ed isolamento dei gruppi

$$DB = \left(\frac{1}{k}\right)\sum_{r=1}^{k}\left[\max_{s=1,2,\cdots,k;s\neq r}\{Q_{r,s}\}\right] \quad dove \quad Q_{r,s} = \left[\frac{\sqrt{\Psi_r}+\sqrt{\Psi_s}}{d(\mu_r,\mu_s)}\right]$$

inoltre

$$\Psi_r = \sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_{ir}\gamma_{jr}d_{ij},\ \Psi_s = \sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_{is}\gamma_{jr}d_{ij} \quad \text{dispersione nei cluster}$$

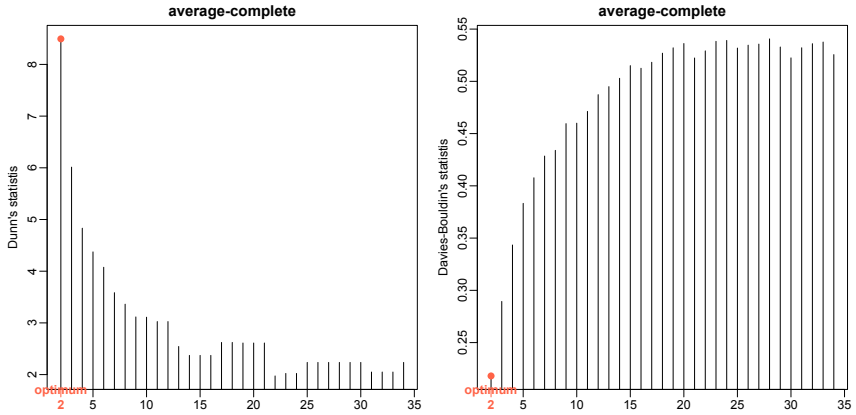
$$d(\mu_r,\mu_s) = \text{distanza tra i centroidi dei de cluster}$$

Dubes (1982) raccomanda questa strategia: se la sequenza dell'indice Davies-Bouldin ha un minimo molto significativo questo è un buon candidato per il numero di cluster.

Se il minimo è a k=2 questo sarà accettabile solo se si può notare una forte caduta dei valori dell'indice per k>2.

Anche per questo indice esistono diverse scelte per isolamento e compatteza dei cluster

Esempio (ancora sparrows data)



Si conferma che la suddivione in due gruppi è la soluzione cercata dalla clustering gerarchica anche se non è la più adatta

Gap statistic

E' stata proposta da Tibshirami et al. (2001). Cerca il numero ottimale di cluster con un confronto tra la clustering ottenuta con k gruppi e quella ottenibile da un data set ottenuto con delle repliche simulate del data set.

Misuriamo l'omogoneità del cluster C_r con l'espressione

$$h_r = \sum_{i=1}^{n}\sum_{j=1}^{n}d_{ij}\gamma_{ir}\gamma_{jr};\ \gamma_{ir} = \begin{cases}1 & U_i \in C_r\\ 0 & U_i \in C\end{cases};\ \gamma_{jr} = \begin{cases}1 & U_j \in C_r\\ 0 & U_j \in C\end{cases}$$

Il prodotto degli indicatori γ è uno se entrambe le unità appartengono a C_r

La qualità della clustering in k gruppi è misurata da

$$H_k = \sum_{r=1}^{k} w_r h_r \quad con \quad w_r = \begin{cases}\left(\frac{1}{2n_r}\right) & se\ \ n_r > 0\\ 0 & se\ \ n_r > 0\end{cases}$$

Valori piccoli di H_k indicano compattezza nei cluster.

Gap statistic/2

Se le dissimilarità fossero delle distanze euclidee allora H_k coinciderebbe con la somma dei quadrati degli scarti dalle rispettive medie nei vari cluster

In questo caso H_k diminuirebbe con k e sarebbe poco informativo dato che per ottenere un miglioramento basterebbe solo aumentare k.

Inoltre, la presenza di gruppi ben strutturati provocherebbe una discesa netta di H_k in corrispondenza del numero più plausibile di gruppi perché le unità si avvicinerebbero maggiormente ai loro centri.

Poiché l'euclideità della matrice delle distanze non è garantita dobbiamo trovare una alterativa

Misuriamo l'omogoneità di una clustering in k gruppi C_r con

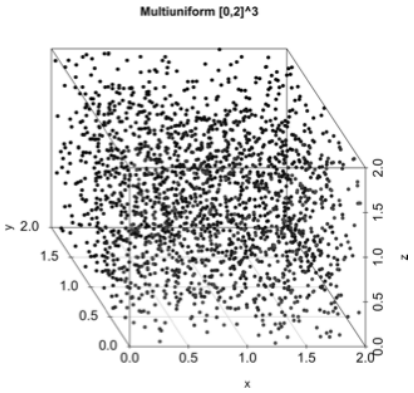
$$W_k = \log(H_k)$$

Gap statistic/2

Si confronta la clustering ottenuta con quella che si avrebbe se la matrice dei dati fosse formata da valori casuali privi di struttura di gruppo. Ad esempio una distribuzione uniforme mutlivariata.

Per mantenere una certa coerenza con i nostri dati, i valori casuali dovrebbero avere almeno lo stesso campo di variazione e la stessa matrice di correlazione (almeno simile)

	X_1	X_2	$\cdots$	X_m
min	L_1	L_2	$\cdots$	L_m
max	U_1	U_2	$\cdots$	U_m



E' necessario ripetere la procedura un certo numero di volte. Diciamo che 60 repliche potrebbero bastare.

Per ogni replica si calcola W_k

Gap statistic/3

In base alle repliche effettuate possiamo calcolare il valore medio di W_k in caso di dati non strutturati

Il confronto di tale media con il valore osservato nel data set reale si realizza con

$$G_k = \frac{\sum_{r=1}^{N} W_k^*}{N} - W_k$$

N è il numero di repliche. Ad esempko N=60.

Il numero di cluster da valutare come migliore scelta è dato dal criterio

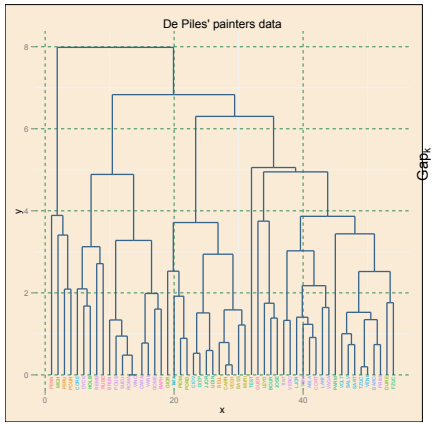
$$k\_opt = \min_{2 \le k \le k'} \left\{ G_k \ge G_{k+1} - S_{k+1}\sqrt{\frac{N+1}{N}} \right\}$$

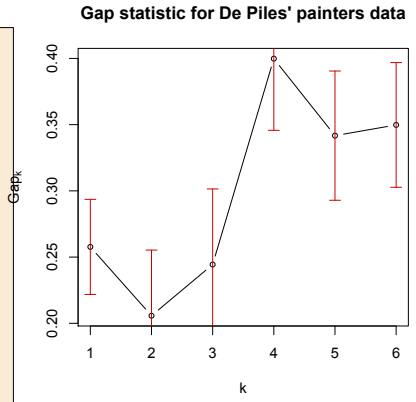
Dove S_k è la deviazione standard di W_k nelle N repliche

clusGap {cluster}

Esempio

E' un data set che valuta alcuni aspetti della tenica di diversi pittori.





K=4 è una scelta ragionevole

Si può usare il pacchetto NbClust

Altro esempio (NbClust)

Gli indici disponibili sono una trentina. Usiamo solo quello discussi nel nostro corso.

Indice	*k*
Dunn	5
Davies – Bouldin	5
Silhouette	5
Gap	5

C'è convergenza su k=5 che si conferma il corretto numero di cluster