Metodologie di Analisi per dati multivariati

L’analisi statistica *multivariata* studia le proprietà di un insieme di *p variabili* rilevate su un insieme di elementi $I = \{I_1, I_2, \ldots\ldots, I_n\}$ *(prodotti, marchi, aziende, individui, …….)*

Matrice di dati multivariati

I dati consistono in una matrice in cui *p* variabili vengono rilevate su *n* di soggetti, oggetti o altre entità di interesse. Tali dati possono essere rappresentati da una matrice *X* ovvero la *matrice dei dati multivariati*

- X = $\begin{bmatrix} x_{11} & \cdot & \cdot & \cdot & x_{1p} \\ \cdot & & & & \cdot \\ x_{i1} & \cdot & \cdot & \cdot & x_{ip} \\ \cdot & & & & \cdot \\ x_{n1} & \cdot & \cdot & \cdot & x_{np} \\ \cdot & & & & \cdot \\ x_{n1} & \cdot & \cdot & \cdot & x_{np} \end{bmatrix}$

Esempio

- Dieta giornaliera 70/80 gr di proteine al giorno suddivise nei diversi gruppi di cibo tra cui quelle che per esercizio si propongono nella prossima tabella
- (libera elaborazione da DASL data base su 25 paesi e diverse variabili)

Esempio

	Country	RedMeat	WhitM	Eggs	Milk	Fish	Frveg
1	Finland	9.5	4.9	2.7	33.7	5.8	1.4
2	France	18	9.9	3.3	19.5	5.7	6.5
3	Italy	9	5.1	2.9	13.7	3.4	6.7
4	Spain	7.1	3.4	3.1	8.6	7	7.2
5	Sweden	9.9	7.8	3.5	24.7	7.5	2
6	UK	17.4	5.7	4.7	20.6	4.3	3.3

Esplorazione iniziale di dati multivariati

Data una matrice dei dati X (nxp) o Z (nxp) (dopo l'eventuale standardizzazione)

- Quali analisi si potrebbero fare?

- Analisi grafica esplorativa
- Analisi tramite statistiche univariate
- Analisi delle correlazioni

La procedura CORR

6 Variabili:	RedMeat WhitM Eggs Milk Fish Frveg

Statistiche semplici

Variabile	N	Media	Dev std	Somma	Minimo	Massimo
RedMeat	6	11.81667	4.66108	70.90000	7.10000	18.00000
WhitM	6	6.13333	2.33295	36.80000	3.40000	9.90000
Eggs	6	3.36667	0.71181	20.20000	2.70000	4.70000
Milk	6	20.13333	8.71703	120.80000	8.60000	33.70000
Fish	6	5.61667	1.55874	33.70000	3.40000	7.50000
Frveg	6	4.51667	2.58567	27.10000	1.40000	7.20000

Coefficienti di correlazione di Pearson, N = 6
Prob > |r| sotto H0: Rho=0

	RedMeat	WhitM	Eggs	Milk	Fish	Frveg
RedMeat	1.00000	0.66832 0.1468	0.65666 0.1566	0.15272 0.7727	-0.30698 0.5540	-0.02143 0.9679
WhitM	0.66832 0.1468	1.00000	0.17182 0.7448	0.26026 0.6184	0.12686 0.8107	-0.06443 0.9035
Eggs	0.65666 0.1566	0.17182 0.7448	1.00000	-0.05458 0.9182	-0.17425 0.7413	-0.18980 0.7187
Milk	0.15272 0.7727	0.26026 0.6184	-0.05458 0.9182	1.00000	0.12845 0.8084	-0.89260 0.0167
Fish	-0.30698 0.5540	0.12686 0.8107	-0.17425 0.7413	0.12845 0.8084	1.00000	-0.23182 0.6585
Frveg	-0.02143 0.9679	-0.06443 0.9035	-0.18980 0.7187	-0.89260 0.0167	-0.23182 0.6585	1.00000

La procedura UNIVARIATE
Variabile: RedMeat

Momenti			
N	6	Somma dei pesi	6
Media	11.8166667	Somma delle osservazioni	70.9
Deviazione std	4.66107999	Varianza	21.7256667
Skewness	0.78125129	Curtosi	-1.7508067
SS non corretta	946.43	SS corretta	108.628333
Coeff var	39.4449646	Errore std media	1.90287794

Misure statistiche di base			
Posizione		Variabilità	
Media	11.81667	Deviazione std	4.66108
Mediana	9.70000	Varianza	21.72567
Moda	.	Range	10.90000
		Range interquartile	8.40000

Test di posizione: Mu0=0				
Test	Statistica		P-value	
T di Student	t	6.209892	Pr > |t|	0.0016
Segno	M	3	Pr >= |M|	0.0313
Rango con segno	S	10.5	Pr >= |S|	0.0313

La procedura UNIVARIATE
Variabile: WhitM

Momenti			
N	6	Somma dei pesi	6
Media	6.13333333	Somma delle osservazioni	36.8
Deviazione std	2.33295235	Varianza	5.44266667
Skewness	0.8172119	Curtosi	0.09046015
SS non corretta	252.92	SS corretta	27.2133333
Coeff var	38.0372666	Errore std media	0.95242381

Misure statistiche di base			
Posizione		Variabilità	
Media	6.133333	Deviazione std	2.33295
Mediana	5.400000	Varianza	5.44267
Moda	.	Range	6.50000
		Range interquartile	2.90000

Test di posizione: Mu0=0				
Test	Statistica		P-value	
T di Student	t	6.43971	Pr > |t|	0.0013
Segno	M	3	Pr >= |M|	0.0313
Rango con segno	S	10.5	Pr >= |S|	0.0313

La procedura UNIVARIATE
Variabile: Eggs

Momenti			
N	6	Somma dei pesi	6
Media	3.36666667	Somma delle osservazioni	20.2
Deviazione std	0.71180522	Varianza	0.50666667
Skewness	1.62670167	Curtosi	3.05069252
SS non corretta	70.54	SS corretta	2.53333333
Coeff var	21.1427292	Errore std media	0.29059326

Misure statistiche di base			
Posizione		Variabilità	
Media	3.366667	Deviazione std	0.71181
Mediana	3.200000	Varianza	0.50667
Moda	.	Range	2.00000
		Range interquartile	0.60000

Test di posizione: Mu0=0				
Test	Statistica		P-value	
T di Student	t	11.58549	Pr > |t|	<.0001
Segno	M	3	Pr >= |M|	0.0313
Rango con segno	S	10.5	Pr >= |S|	0.0313

La procedura UNIVARIATE
Variabile: Milk

Momenti			
N	6	Somma dei pesi	6
Media	20.1333333	Somma delle osservazioni	120.8
Deviazione std	8.71703313	Varianza	75.9866667
Skewness	0.3585663	Curtosi	0.26240993
SS non corretta	2812.04	SS corretta	379.933333
Coeff var	43.2965222	Errore std media	3.55871388

Misure statistiche di base			
Posizione		Variabilità	
Media	20.13333	Deviazione std	8.71703
Mediana	20.05000	Varianza	75.98667
Moda	.	Range	25.10000
		Range interquartile	11.00000

Test di posizione: Mu0=0				
Test	Statistica		P-value	
T di Student	t	5.657475	Pr > |t|	0.0024
Segno	M	3	Pr >= |M|	0.0313
Rango con segno	S	10.5	Pr >= |S|	0.0313

La procedura UNIVARIATE
Variabile: Fish

Momenti			
N	6	Somma dei pesi	6
Media	5.61666667	Somma delle osservazioni	33.7
Deviazione std	1.55873881	Varianza	2.42966667
Skewness	-0.3042165	Curtosi	-1.1045526
SS non corretta	201.43	SS corretta	12.1483333
Coeff var	27.7520262	Errore std media	0.63635245

Misure statistiche di base			
Posizione		Variabilità	
Media	5.616667	Deviazione std	1.55874
Mediana	5.750000	Varianza	2.42967
Moda	.	Range	4.10000
		Range interquartile	2.70000

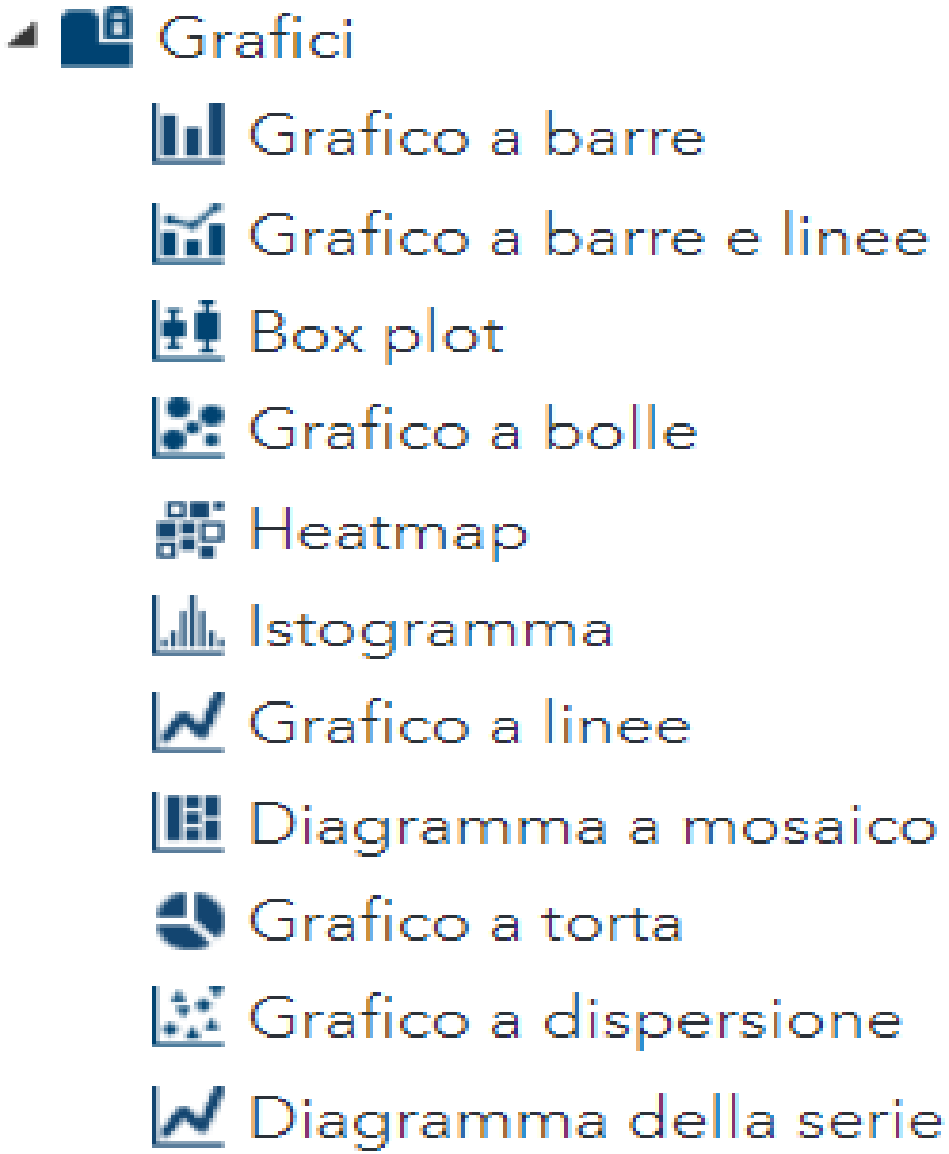
Test di posizione: Mu0=0				
Test	Statistica		P-value	
T di Student	t	8.826346	Pr > |t|	0.0003
Segno	M	3	Pr >= |M|	0.0313
Rango con segno	S	10.5	Pr >= |S|	0.0313

La procedura UNIVARIATE
Variabile: Frveg

Momenti			
N	6	Somma dei pesi	6
Media	4.51666667	Somma delle osservazioni	27.1
Deviazione std	2.58566561	Varianza	6.68566667
Skewness	-0.181949	Curtosi	-2.6997223
SS non corretta	155.83	SS corretta	33.4283333
Coeff var	57.2472092	Errore std media	1.05559357

Misure statistiche di base			
Posizione		Variabilità	
Media	4.516667	Deviazione std	2.58567
Mediana	4.900000	Varianza	6.68567
Moda	.	Range	5.80000
		Range interquartile	4.70000

Test di posizione: Mu0=0				
Test	Statistica		P-value	
T di Student	t	4.278793	Pr > |t|	0.0079
Segno	M	3	Pr >= |M|	0.0313
Rango con segno	S	10.5	Pr >= |S|	0.0313


Grafici
Grafico a barre
Grafico a barre e linee
Box plot
Grafico a bolle
Heatmap
Istogramma
Grafico a linee
Diagramma a mosaico
Grafico a torta
Grafico a dispersione
Diagramma della serie

La segmentazione del database

- La segmentazione del database ha come obiettivo la suddivisione di un database in segmenti che contengono dei records che hanno in comune un determinato numero di proprietà, e che per questo motivo vengono considerati fra di loro omogenei.

La segmentazione del database

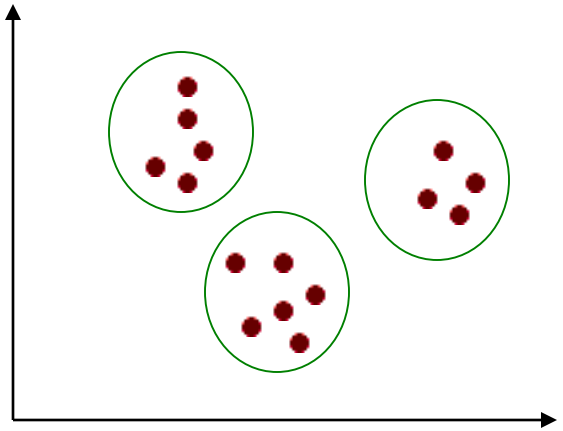
- *Segmentazione* descrive l'operazione di data mining,
  - *segmenti* descrivono i gruppi risultanti dai records contenuti nel database.
  - I segmenti devono essere caratterizzati da un'elevata omogeneità interna (internamente al segmento) e da un'elevata eterogeneità esterna (fra i segmenti).
  - Per omogeneità interna intendiamo che i records sono il più possibile "vicini" l'uno all'altro.

Cluster Analysis

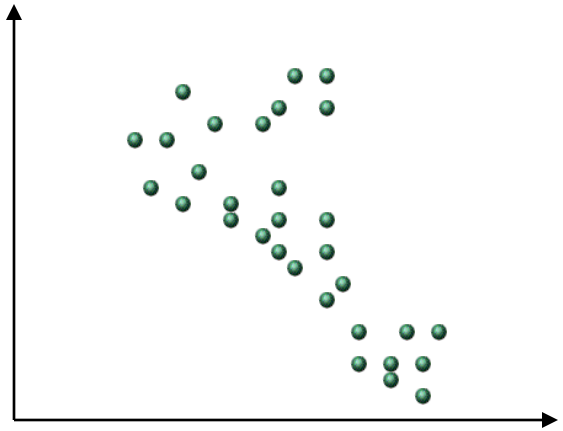
- Una delle tecniche che vengono utilizzate nella operazione di segmentazione del database è la *cluster analysis*.
- Obiettivo della cluster analysis è individuare all'interno di un dataset gruppi di elementi raggruppati in base al valore assunto da una misura di distanza\similarità misurata su alcune variabili rilevanti.

Obiettivo

- Gruppi ben definiti



- Gruppi non ben definiti



- Gli algoritmi di cluster analysis presuppongono la scelta di una misura tramite la quale sia possibile valutare l'omogeneità tra gli "elementi". Più specificatamente, l'omogeneità viene definita in termini di minore *distanza* quando il dataset è costituito esclusivamente da dati quantitativi, mentre si parlerà di *maggiore similarità* in presenza di variabili qualitative (attributi) o miste (qualitative/quantitative).

distanze

- Siano A e B due “elementi” (per esempio due records) individuati rispettivamente dai vettori **x** = (x_1, x_2, ..., x_p) e **y** = (y_1, y_2,...., y_p) relativi a *p* variabili .
- Una funzione reale *d(A,B)* è una *distanza* se gode delle seguenti proprietà:
-
- $d(A,B) = d(B,A)$ (proprietà simmetrica)
- $d(A,B) \geq 0$ (non negatività)
- $d(A,A) = 0$ (identità)

distanze

inoltre se valgono le seguenti proprietà, le funzioni di distanza vengono dette metriche:

- $d(A,B) = 0$ se e solo se $A = B$
- $d(A,B) \leq d(A,C) + d(B,C)$ (disuguaglianza triangolare) dove C è un terzo elemento individuato dal vettore $\mathbf{z} = (z_1, z_2, \ldots, z_p)$.

distanze

- Sia **X** una matrice (*nxp)* $\mathbf{X}=(\mathbf{x}_1 ,\ldots\ldots, \mathbf{x}_n)'$.
- La distanza euclidea al quadrato tra i punti generici $\mathbf{x}_i$ e $\mathbf{x}_j$ è data da

- $d^2_{ij} = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 = || \mathbf{x}_i - \mathbf{x}_j ||^2$ *i,j=1,…,n*

distanze

Sia **X** una matrice (*nxp*) **X**=($\mathbf{x_1}$,……., $\mathbf{x_n}$)’

- Se le variabili utilizzate per la clusterizzazione hanno una diversa unità di misura, è possibile standardizzarle. Questa procedura equivale alla considerare la distanza detta di *Karl Pearson,* tale distanza è invariante per cambiamenti di scala

$$d^2_{ij} = \sum_{k=1}^{p} \frac{(x_{ik} - x_{jk})^2}{\sigma_k^2} \qquad i,j=1,\ldots,n$$

distanze

- Sia la distanza euclidea che la distanza di *Karl Pearson,* possono essere considerate casi particolari della *distanza Minkowski ponderata*

$$d_{ij} = \left\{ \sum_{k=1}^{p} W_k \left| x_{ik} - x_{jk} \right|^l \right\}^{\frac{1}{l}} \quad i,j=1,\ldots,n$$

- per $w_k = 1$ e $l = 2$, $w_k = 1/\sigma_k^2$ e $l = 2$
- w_k peso attribuito alla k-esima variabile

distanze

- Le funzioni distanza che abbiamo definito, non tengono conto direttamente delle interdipendenze che possono esistere tra le p variabili utilizzate . La *distanza di Mahalonobis* permette invece di eliminare questo inconveniente:

$$d_M(\mathbf{x}_i, \mathbf{x}_h) = (\mathbf{x}_i - \mathbf{x}_h)' \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_h) = \sum_{k=1}^{p} \sum_{l=1}^{p} s^{kl} (x_{ik} - x_{hk})(x_{il} - x_{hl})$$

distanze

$$d_M(\mathbf{x}_i, \mathbf{x}_h) = (\mathbf{x}_i - \mathbf{x}_h)' \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_h) = \sum_{k=1}^{p} \sum_{l=1}^{p} s^{kl} (x_{ik} - x_{hk})(x_{il} - x_{hl})$$

- dove: *S* è la matrice di varianza e covarianza (pxp) tra le *p* variabili $x_{k,}$ o matrice delle correlazioni se le variabili sono standardizzate
- s^{kl} è il generico elemento della k-esima riga e i-esima colonna della matrice inversa di *S*,
- $\mathbf{x_i}$ e $\mathbf{x_h}$ sono i vettori colonna (px1) che contengono gli elementi della i-esima e h-esima riga della matrice X che rappresenta il dataset di partenza.

Indici di similarità

- per i dati qualitativi o misti (qualitativi / quantitativi) si applica il concetto di *similarità.*
- Dati due "elementi" A e B caratterizzati da *p* attributi, una misura di *similarità* s(A,B) deve godere delle seguenti proprietà:
- $s(A,B) = s(B,A)$ (simmetria)
- $s(A,B) \geq O$ (non negatività)
- s(A,B) cresce al crescere della similarità fra A e B

Indici di similarità per caratteri dicotomici

- Dati due elementi A e B, individuati rispettivamente dai vettori **x** = (x_1, x_2, ..., x_p) e **y** = (y_1, y_2,...., y_p) relativi a *p* attributi, si utilizzano delle variabili dicotomiche per denotare la presenza o assenza dei *p* attributi:

$$\begin{cases} x_k = 1 & \Rightarrow k\text{ - esimo elemento presente} \\ x_k = 0 & \Rightarrow K\text{ - esimo elemento assente} \end{cases}$$

Indici di similarità per caratteri dicotomici

- Si ponga
- $a = \sum_{k=1}^{p} x_k y_k \quad b = \sum_{k=1}^{p} (1 - x_k)(1 - y_k)$

Dove le quantità *a* e *b* sono rispettivamente le frequenze di $(x_k, y_k) = (1,1)$ e $((1-x_k), (1-y_k)) = (0,0)$.

Indici di similarità per caratteri dicotomici

- La più semplice misura di similarità fra A e B e data dall'indice di similarità di Russel e Rao:

$$s(A,B) = \frac{a}{p}$$

Indici di similarità per caratteri dicotomici

- Un alternativa è data dal coefficiente di *simple matching,* proposto da Sokal e Michener (1958), definito come

$$s_1(A,B) = \frac{a+b}{p}$$

Indici di similarità per caratteri qualitativi e quantitativi

OSS.

- Esistono indici per valutare la prossimità congiunta di caratteri qualitativi e quantitativi come l'indice di Gower (1971)

Esempio

	Country	RedMeat	WhitM	Eggs	Milk	Fish	Frveg
1	Finland	9.5	4.9	2.7	33.7	5.8	1.4
2	France	18	9.9	3.3	19.5	5.7	6.5
3	Italy	9	5.1	2.9	13.7	3.4	6.7
4	Spain	7.1	3.4	3.1	8.6	7	7.2
5	Sweden	9.9	7.8	3.5	24.7	7.5	2
6	UK	17.4	5.7	4.7	20.6	4.3	3.3

	Dist1	Dist2	Dist3	Dist4	Dist5	Dist6
1	0	.	.	.	.	.
2	15.177574104	0	.	.	.	.
3	11.934106472	10.903263141	0	.	.	.
4	14.909707465	15.642756706	6.489855342	0	.	.
5	5.1250074877	8.6273781763	13.902338883	11.792357216	0	.
6	14.609401978	9.4803033268	12.397660842	18.078243003	10.930045857	0

	Country
1	Finland
2	France
3	Italy
4	Spain
5	Sweden
6	UK

risultati su dati leggermente diversi

DISTANZA EUCLIDEA AL QUADRATO

	(1) FIN	(2) FRA	(3) ITA	(4) SPA	(5) SUE	(6) GB
FINLANDIA	0	10.6	11.9	14.5	3.6	14.5
FRANCIA	10.6	0	6.7	7.9	7.8	6.2
ITALIA	11.9	6.7	0	5.9	12.6	12.3
SPAGNA	14.5	7.9	5.9	0	8.2	17.1
SVEZIA	3.6	7.8	12.6	8.2	0	10.1
GRAN BRET.	14.4	6.2	12.3	17.1	10.1	0

$$d^2_{i2} = \sum_{k=1}^{p} (z_{1k} - z_{2k})^2 = \| \underline{z}_1 - \underline{z}_2 \|^2$$

$$Z = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & & & \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{pmatrix} \begin{matrix} \underline{z}_1' \\ \underline{z}_2' \\ \\ \underline{z}_n' \end{matrix}$$

Fasi del processo di analisi dei clusters

- La *cluster analysis* tradizionale consiste in alcune fasi strettamente collegate tra loro, la procedura può richiedere una serie di tentativi e di ripetizioni dei vari passaggi che vengono di seguito sintetizzati:
- scelta delle unità di osservazione
- Scelta delle variabili;
- Scelta della misura di similarità/distanza tra unità statistiche;
- scelta dell'algoritmo di *clustering*;
- validazione dei *clusters*;
- interpretazione dei risultati.

Fasi del processo di analisi dei clusters in DM

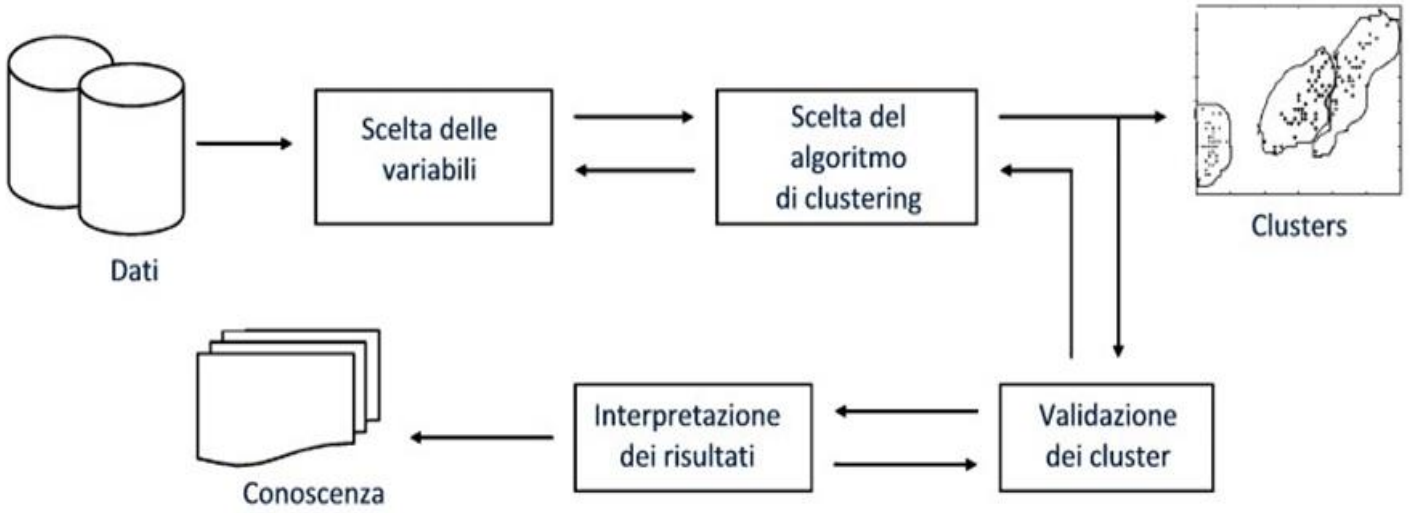


Figura 1.1- Le fasi di una procedura di clustering
Fonte: Xu, Wunsch (2005)

Gli algoritmi di cluster analysis:

- Esistono due grandi famiglie di classificazione.
- *Gli algoritmi di classificazione gerarchica,* caratterizzati da una procedura iterativa che genera una gerarchia nelle partizioni
- *Gli algoritmi di classificazione diretta,* che si basano per la formazione delle classi sulla minimizzazione di una funzione obiettivo assegnata.

- Metodi gerarchici
  - Agglomerativi
  - Divisivi
- Metodi non gerarchici
  - Metodi basati sull’errore quadratico
  - Metodi basati sui modelli mistura
  - Metodi basati sulla teoria dei grafi
  - Altri metodi

Gli algoritmi di cluster analysis:
Gli algoritmi di classificazione gerarchica

- *I metodi gerarchici* sono procedure che si sviluppano per fasi ordinate, in modo che in ogni fase vengano aggregati i due “elementi” (o gruppi di elementi) che risultano più omogenei tra loro, in base alla misura di distanza (similarità) prescelta.
- Una proprietà peculiare di tali procedure è data dal fatto che nel momento in cui un oggetto viene allocato in un gruppo, non può venire riallocato in una fase successiva in un diverso cluster.

Gli algoritmi di cluster analysis:

Gli algoritmi di classificazione gerarchica

- *I metodi gerarchici* sono procedure che si sviluppano per fasi ordinate, in modo che in ogni fase vengano aggregati i due "elementi" (o gruppi di elementi) che risultano più omogenei tra loro, in base alla misura di distanza (similarità) prescelta.

Esistono due grandi gruppi di metodi gerarchici:

- *agglomerativi,* in cui i clusters sono formati raggruppando i casi in clusters via via più grandi, fino a che tutte le osservazioni sono raggruppate in un unico cluster;
- *divisivi,* in cui invece si parte con i clusters raggruppati in un singolo cluster per giungere ad ottenere tanti clusters quante sono le osservazioni.

Gli algoritmi di cluster analysis:

Gli algoritmi di classificazione gerarchica

Data una matrice dei dati X (nxp) o Z (nxp) (dopo l'eventuale standardizzazione)

- sia definita una misura di distanza (similarità) fra le coppie di "elementi" che si desidera classificare e venga costruita una matrice ***D*** di dimensioni (nxn) che contenga le distanze (similarità) tra tutte le coppie di "elementi";

$$\mathbf{D} = \begin{bmatrix} 0 & d_{12} & . & . & d_{1n} \\ d_{21} & 0 & . & . & d_{2n} \\ . & . & . & . & . \\ . & . & . & . & . \\ d_{n1} & . & . & d_{n,n-1} & 0 \end{bmatrix}$$

Gli algoritmi di classificazione gerarchica

- Siano definite delle regole per decidere quali "elementi" o quali raggruppamenti di elementi debbano essere riuniti in ogni fase dell'algoritmo.Più specificatamente, lo sviluppo di una procedura gerarchica segue sempre alcune fasi tipiche.
- FASE 0. E' lo stadio iniziale, in cui si considerano gli *n* "elementi" che indichiamo con e_1, $e_{2, \ldots\ldots\ldots,}$ e_n come *n* clusters elementari $\{e_1\}$, $\{e_2\}$,..., $\{e_n\}$.

Gli algoritmi di classificazione gerarchica

- FASE 1. Si sceglie, osservando la matrice delle distanze ***D***, l'elemento d_{ij} *per* cui si ha

$$d_{ij} = \min_{hl}(d_{hl})$$

h=1,…….,n l=1 ,……, n

- I due clusters elementari $\{e_i\}$ e $\{e_j\}$ corrispondenti vengono aggregati ottenendo così il primo cluster "composito" $\{e_i, e_j\}$.
- A conclusione di questa fase si hanno pertanto (*n-2*) clusters elementari $\{e_1\}$, …..$\{e_{n-2}\}$ ed il cluster composito $\{e_i, e_j\}$.

Gli algoritmi di classificazione gerarchica

- FASE 2. Viene riaggiornata la matrice delle distanze ***D,*** per tenere conto dell'aggregazione avvenuta, eliminando le righe e le colonne relative agli elementi e_i, e_j, che sono stati riuniti in un unico cluste*r* e viene aggiunta una riga e una colonna contenente le distanze tra il cluster composito $\{e_i, e_j\}$ e gli (n-2) clusters rimanenti.

Gli algoritmi di classificazione gerarchica

- FASI SUCESSIVE (da 3 a n-1). L'algoritmo ripete le fasi 1 e 2 fino a quando gli n elementi non vengono raggruppati in un unico cluster, quindi il processo ha termine dopo (n – 1) raggruppamenti.

Gli algoritmi di classificazione gerarchica

- Il punto chiave dell'analisi è proprio la definizione di distanza/similarità tra un cluster composito creatosi in una delle *n*-1 fasi e i rimanenti "elementi"(o clusters compositi).
- Esistono, per quanto concerne questo problema, diverse soluzioni possibili che definiscono altrettante varianti degli algoritmi gerarchici.

- METODO DEL LEGAME SEMPLICE. Sia C_r un cluster contenente n_r elementi e_i e C_s un cluster contenente n_s elementi e_j. La distanza tra C_s e C_r è data da

$$d_{C_r,C_s} = \min_{e_i \in C_r, e_j \in C_s} (d_{ij})$$

dove d_{ij} e la distanza tra il generico elemento $e_i \in C_r$ e il generico elemento $e_j \in C_s$.

risultati su dati leggermente diversi

DISTANZA EUCLIDEA AL QUADRATO

	(1) FIN	(2) FRA	(3) ITA	(4) SPA	(5) SUE	(6) GB
FINLANDIA	0	10.6	11.9	14.5	3.6	14.5
FRANCIA	10.6	0	6.7	7.9	7.8	6.2
ITALIA	11.9	6.7	0	5.9	12.6	12.3
SPAGNA	14.5	7.9	5.9	0	8.2	17.1
SVEZIA	3.6	7.8	12.6	8.2	0	10.1
GRAN BRET.	14.4	6.2	12.3	17.1	10.1	0

$$d^2_{i2} = \sum_{k=1}^{p} (z_{1k} - z_{2k})^2 = \| \underline{z}_1 - \underline{z}_2 \|^2$$

$$Z = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & & & \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{pmatrix} \begin{matrix} \underline{z}_1' \\ \underline{z}_2' \\ \\ \underline{z}_n' \end{matrix}$$

LEGAME SINGOLO

(1;5)	2	3	4	6	
0	7.8	11.8	8.2	10.1	(1;5)
	0	6.7	7.8	6.2	2
		0	5.8	12.3	3
			0	17.1	4
				0	6

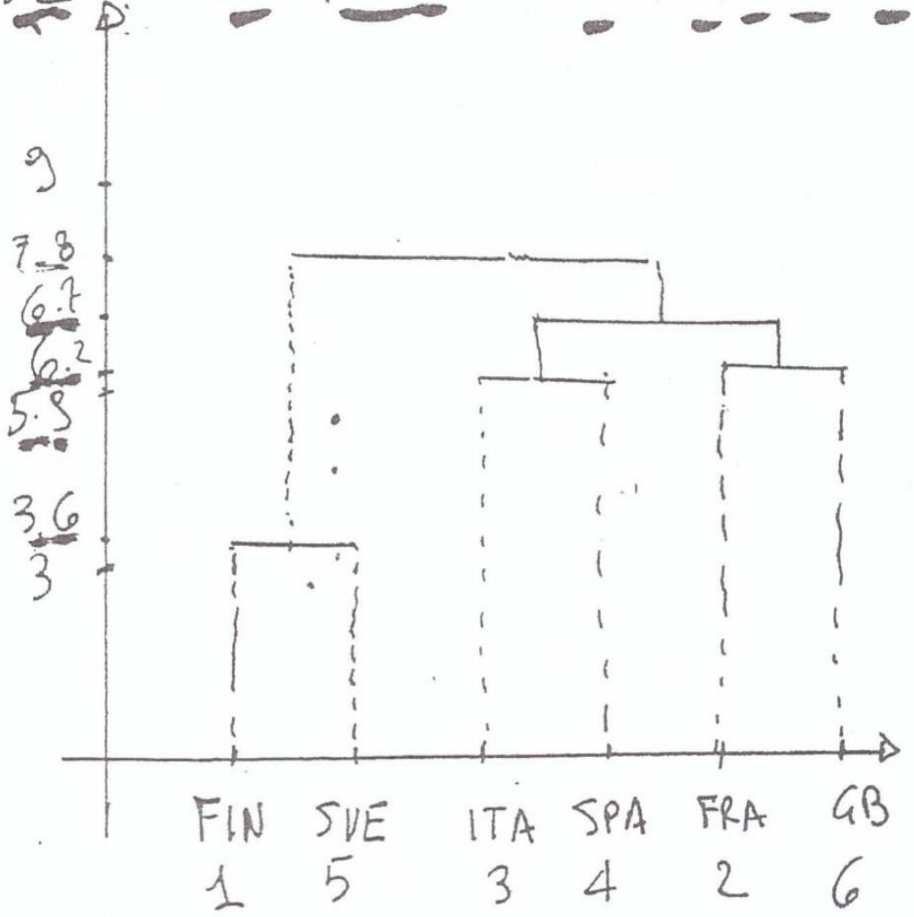
$d_{C_r C_s} = \min\{d_{ij} \mid l_i \in C_r,\ l_j \in C_s\}$

(1;5)	2	(3;4)	6	
0	7.8	8.2	10.1	(1;5)
	0	6.7	6.2	2
		0	12.3	(3;4)
			0	6

(1;5)	(2;6)	(3;4)	
0	7.8	8.2	(1;5)
	0	6.7	(2;6)
		0	(3;4)

(1;5)	(2;3;4;6)	
0	7.8	(1;5)
	0	

DENDROGRAMMA UTILIZZANDO IL LEGAME SEMPLICE



- METODO DEL LEGAME COMPLETO. Sia C_r un cluster contenente n_r elementi e_i, e C_s un cluster contenente n_s elementi e_j. La distanza tra C_s e C_r è data da
- $$d_{C_r,C_s} = \max_{e_i \in C_r, e_j \in C_s} (d_{ij})$$

- dove d_{ij} e la distanza tra il generico elemento $e_i \in C_r$ e il generico elemento $e_j \in C_s$

LEGAME COMPLETO

(1;5)	2	3	4	6	
0	10.6	12.6	14.5	14.4	(1;5)
	0	6.7	7.8	6.2	2
		0	5.8	12.3	3
			0	17.1	4
				0	6

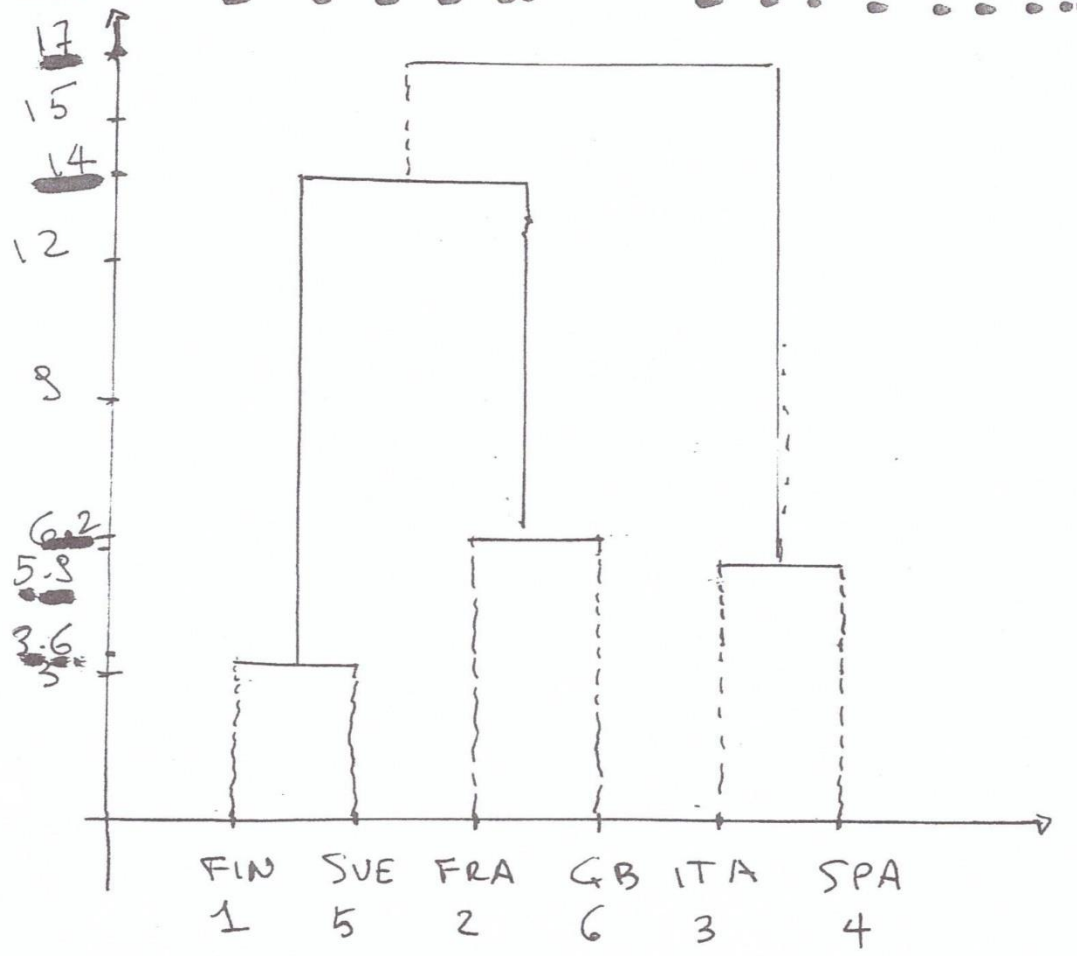
$$d_{C_r C_s} = \max_{\substack{e_i \in C_r \\ e_j \in C_s}} \{ d_{ij} \}$$

(1;5)	2	(3;4)	6	
0	10.6	14.5	14.4	(1;5)
	0	7.8	6.2	2
		0	17.1	(3;4)
			0	6

(1;5)	(2;6)	(3;4)	
0	14.4	14.5	(1;5)
	0	17.1	(2;6)
		0	(3;4)

(1;2;5;6)	(3;4)	
0	17.1	(1;2;5;6)
	0	

DENDROGRAMMA UTILIZZANDO IL LEGAME COMPLETO



- METODO DEL LEGAME MEDIO FRA I GRUPPI. Dati due clusters C_r e C_s contenenti rispettivamente n_r elementi e_i, e n_s elementi e_j. L'omogeneità tra C_s e C_r è misurata con la media aritmetica dei valori della distanza tra ogni elemento del cluster C_r e del cluster C_s.

$$d_{c_r c_s} = \frac{1}{n_r n_s} \sum_{e_i \in c_r} \sum_{e_j \in c_s} d_{ij}$$

- dove d_{ij} e la distanza tra il generico elemento $e_i \in C_r$ e il generico elemento $e_j \in C_s$.

- Dopo il primo step in cui si sono uniti FIN e SVE

la distanza tra il cluster composito {FIN, SVE} e per es. la Francia {FRA} è

1/(2*1) [(10.6+7.8)/2]= 9.2

$$d_{c_r c_s} = \frac{1}{n_r n_s} \sum_{e_i \in c_{r\,s}} \sum_{ej \in c_s} d_{ij}$$

- METODO DEL CENTROIDE. La distanza tra due clusters è data dalla distanza dei *centri* dei clusters stessi; dove i centri possono essere definiti come valore medio di ciascuna variabile, calcolato su tutte le unita appartenenti ad un cluster.

$$d_{c_r,c_s} = \left\| \overline{\boldsymbol{x}_r} - \overline{\boldsymbol{x}_s} \right\| = \left(\sum_{k=1}^{p} \left(\overline{x_{rk}} - \overline{x_{sk}} \right)^2 \right)^{\frac{1}{2}}$$

- Dopo il primo step in cui si sono uniti FIN e SVE

distanza tra il cluster composito {FIN, SVE} e la Francia {FRA}

Centroide FRA

2	France	18	9.9	3.3	19.5	5.7	6.5

centroide {FIN, SVE}

1	Finland	9.5	4.9	2.7	33.7	5.8	1.4
5	Sweden	9.9	7.8	3.5	24.7	7.5	2

(9.7 6.35 3.1 29.2 6.65 1.7)

A questo punto si calcola la distanza tra i centroidi (SAS utilizza la distanza euclidea al quadrato)

- METODO DI WARD e’ diretto alla minimizzazione della varianza all’interno dei gruppi. La conseguenza è che i clusters ottenuti sono il più possibile omogenei, presentando la minima dispersione interna, e contemporaneamente sono il più possibile diversi l’uno dall’altro essendo massimizzata la dispersione tra i clusters.
- Ad ogni passo questo algoritmo tende ad ottimizzare la partizione ottenuta tramite l’aggregazione di due elementi.

METODO DI WARD

- Una partizione si considera tanto migliore quanto piu' le classi risultano omogenee al loro interno e differenti l'una dall'altra. In altri termini, quanto più e' elevata la varianza tra le classi, e bassa la varianza interna (alle classi).
- E' noto che la varianza totale di un insieme di unità, si puo' scomporre nella somma di due quantità: varianza interna (ai cluster) e varianza esterna (cioe' tra i cluster).

METODO DI WARD

In maniera analoga si scompone la matrice di varianze e covarianze S

- S = SW + SB

Dove S e' la matrice di varianze e covarianze totali;

SW e' la matrice delle varianze e covarianze "interne";

SB e' la matrice delle varianze e covarianze "esterne".

L'algoritmo ricerca "il salto minimo di aumento della varianza interna", cioe' ad ogni passo aggrega ad un cluster già individuato, l'unità o il cluster che portino il minor incremento di varianza interna.