La Cluster analysis

(S. Terzi)

1.Introduzione

La cluster analysis e' una tecnica di analisi multivariata attraverso la quale e' possibile raggruppare le unità statistiche, in modo da minimizzare la "lontananza logica" interna a ciascun gruppo e di massimizzare quella tra i gruppi.

La "lontananza logica" viene quantificata per mezzo di misure di similarità/dissimilarità definite tra le unità statistiche.

Date le seguenti proprietà:

1. separabilità $d(i,h) = 0$ se e solo se $x_i = x_h$.
2. simmetria $d(i,h) = d(h,i)$
3. disuguaglianza triangolare:
$$d(i,h) \leq d(i,e) + d(e,h) \quad \forall i, e, h$$
4. condizione di Krassner:
$$d(i,h) \leq \sup(d(i,e); d(e,h)) \; \forall i, e, h$$

Un'applicazione d che associa un valore positivo o nullo a ciascuna coppia (i,h) si definisce:

a) indice di dissimilarità se soddisfa le proprietà 1 e 2;
b) metrica o distanza se soddisfa 1,2 e 3;
c) ultrametrica se soddisfa 1,2 e 4.

La scelta tra indici di dissimilarità e metrica e' legata al tipo di dati che si hanno a disposizione.
Per dati di tipo numerico (quantitativi) possiamo utilizzare delle misure di distanza, ovvero delle metriche.
Per dati di tipo qualitativo bisogna utilizzare misure *matching-type*, cioe' di associazione (similarità o dissimilarità).

Come si effettua una cluster analysis?

Si parte dalla matrice dei dati X di dimensione *nxp* e la si trasforma in una matrice *nxn* di dissimilarità o di distanze tra le n coppie di osservazioni (vettori di p elementi).
Si sceglie poi un algoritmo che definisca le regole su come raggruppare le unità in sottogruppi sulla base delle loro similarità.
Lo scopo e' di identificare un minor numero di gruppi tali che gli elementi appartenenti ad un gruppo siano – in qualche senso – piu' simili tra loro che non agli elementi appartenenti ad altri gruppi.
Il punto di partenza fondamentale e' la definizione di una misura di similarità o di distanza tra gli oggetti (cioe' tra le righe della matrice dei dati).
L'altro punto fondamentale e' la regola in base alla quale si formano i gruppi.

A seconda del tipo di dati, si hanno misure diverse. Per dati quantitativi si hanno misure di distanza; per dati qualitativi si hanno misure di associazione.

1.1.Misure di distanza

Partendo dalla matrice dei dati, ricordiamo che i suoi (n) vettori riga rappresentano le n unità statistiche. Ciascuna unità statistica e' quindi un vettore di p-elementi, contenenti i valori da essa assunti sulla prima, la seconda, la j-esima e la p-esima variabile.
Supponiamo che tali valori siano numeri e non attributi, ovvero supponiamo che le p variabili siano quantitative.
Possiamo definire la distanza tra due unità statistiche, i ed h, in diversi modi.
Utilizzando la metrica di Manhattan:

$$d(i,h) = \Sigma_j \left| x_{ij} - x_{hj} \right|$$

Utilizzando la metrica euclidea:

$$d(i,h) = (\Sigma_j (x_{ij} - x_{hj})^2)^{1/2}$$

Un'altra distanza che a volte viene utilizzata e' la distanza di Mahalanobis (definita in termini vettoriali):

$$d(i,h) = (x_i - x_h)' S^{-1} (x_i - x_h)$$

1.2.Misure di associazione

Si utilizzano per caratteri espressi in scala nominale.
Supponiamo di avere p attributi, ciascuno dei quali puo' essere presente o assente in una generica unità statistica.
Supponiamo di voler confrontare la similarità tra una unità statistica x_i, ed un'altra x_h.
Le righe i-esima ed h-esima della matrice dei dati si presenteranno piu' o meno cosi':

	ATTRIBUTI					
	1	2	…	j	…	p
i	0	1	1	0	1	1
h	1	0	1	0	0	1

(1 indica presenza, 0 indica assenza dell'attributo j).
E potranno generare la seguente matrice di associazione:

		individuo i		tot.
		+	-	
h	+	2	1	3
	-	2	1	3
	tot.	4	2	6

Possiamo definire diversi tipi di coefficienti di similarità.

Posta una tabella:

	+	-
+	a	b
-	c	d

potremmo definire come coefficienti di similarità:

(i) (a+d)/(a+b+c+d);
(ii) a/(a+b+c);
(iii) a/(a+b+c+d)

i quali differiscono tra loro per il modo in cui tengono conto delle associazioni (0,0).
Ma anche:

(iv) (2 a)/(2 a + b+c)
(v) 2(a+d)/(2 (a+d) + b+c)
(vi) a/(a+2(b+c))

che differiscono per il fatto che le “associazioni” hanno peso doppio delle dissociazioni o viceversa, queste ultime pesano il doppio delle prime.

Tali coefficienti sono compresi tra 0 ed 1.
Volendo definire una misura di dissimilarità basterà fare il complemento a 1 dell’indice di similarità prescelto. Per esempio:

$$1 - (a+d)/(a+b+c+d) = b+c/(a+b+c+d)$$

e’ uno degli indici di dissimilarità piu’ comunemente utilizzati.

Le proprietà di un indice di dissimilarità sono:
a)$d(i,i) = 0$
b)$d(i,h) = d(h,i)$.

Ma queste due proprietà da sole non bastano ad evitare che possano verificarsi delle incoerenze. Ad esempio, puo' accadere che, pur essendo $d(i,m) = 0$, risulti $d(i,h) \neq d(m,h)$.
Tale tipo di incoerenza scompare qualora sia verificata almeno una delle seguenti relazioni:
c)$d(i,h) = 0$ solo se $i=h$;
d) $d(i,h) \leq d(i,m) + d(m,h) \quad \forall i, m, h.$

2.Metodi di classificazione

Effettuata la scelta della misura di diversità da utilizzare, si pone la scelta del metodo o algoritmo di classificazione e dell'eventuale criterio di aggregazione/suddivisione.

I metodi di classificazione piu' comuni sono:

1. Metodi gerarchici aggregativi
2. Metodi gerarchici divisivi
3. Metodi non gerarchici.

I **metodi gerarchici** realizzano fusioni o divisioni successive dei dati. Nel caso dei metodi aggregativi (o "agglomerativi") gli n oggetti iniziali vengono fusi in gruppi via via piu' ampi (alla fine: un unico gruppo); nel caso dei metodi divisivi (o "scissori") vengono definite partizioni sempre piu' fini dell'insieme iniziale (alla fine n clusters contenenti ciascuno un elemento). La caratteristica principale che li distingue dai metodi non gerarchici e' che la assegnazione di un oggetto ad un cluster e' **irrevocabile**. Ovvero, una volta che un oggetto e' entrato a far parte di un cluster, non ne viene piu' rimosso.

I **metodi non gerarchici** sono solo di tipo aggregativo, e producono un'unica partizione. Procedono a riallocazioni successive delle unità tra i gruppi definiti a priori, fino alla partizione giudicata "ottima" sulla base di un criterio predefinito.

3. Metodi gerarchici aggregativi

Si suppone che l’insieme di oggetti da classificare sia dotato di una misura di dissimilarità. Immaginiamo per semplicità che si tratti di una distanza.
Si costruisce una prima matrice di distanze fra le n unità statistiche. Si aggregano le due unità piu’ vicine (ovvero con distanza minima), in un cluster.
Al passo successivo una terza unità entra a far parte del cluster trovato al passo precedente, oppure, due unità vengono fuse per formare un diverso cluster.
Si continua a procedere in questo modo finche’ non si sia formato un unico cluster contenente tutte le unità.
Tutto il procedimento poggia sulla definizione del criterio di assegnazione delle unità ai cluster (o di un cluster piccolo ad uno piu’ grande).
Esistono diversi possibili criteri, e conseguentemente, diversi algoritmi aggregativi.
Ne vedremo alcuni:

- **Legame singolo**
- **Legame completo**
- **Legame medio**
- **Metodo del centroide.**

Il metodo del **legame singolo** si basa su un criterio di distanza minima. Supponendo di avere 4 unità : A,B,C,D, e di aver definito un coefficiente di dissimilarità o una misura di distanza tra le unità (d_{AB}, d_{AC}, ……,d_{CD}); supponendo che le unità A e B vengano fuse in un solo cluster, la distanza tra il cluster (AB) e l'unità C e' definita come:

$$d_{(A,B)C} = \min(d_{AC}, d_{BC})$$

Posto che le unità C e D vengano fuse nel cluster (CD), la distanza tra il cluster (AB) ed il cluster (CD) viene definita come:

$$d_{(AB)(CD)} = \min(d_{AC}, d_{AD}\ d_{BC}, d_{BD})$$

Al primo passo si fondono le 2 unità aventi distanza minore, ottenendo cosi' n-1 gruppi.
Si calcola una nuova matrice di distanze tra gli n-1 clusters. Si aggregano i due cluster aventi distanza minima.
E cosi' via, fino ad avere un unico cluster contenente n unità.

Esempio:
Supponiamo di avere per 5 unità statistiche la seguente matrice di distanze:

$D^{(1)} =$

	A	B	C	D	E
A	0	1	5	6	8
B	1	0	3	8	7
C	5	3	0	4	6
D	6	8	4	0	2
E	8	7	6	2	0

Nel primo passo vengono fuse le unità A e B, essendo le piu' "vicine" (con $d_{AB} = 1$).
A questo punto bisogna calcolare le distanze tra questo cluster (AB) e le altre unità.
$d_{(A,B)C} = \min(d_{AC}, d_{BC}) = d_{BC} = 3$
$d_{(A,B)D} = \min(d_{AD}, d_{BD}) = d_{AD} = 6$
$d_{(A,B)E} = \min(d_{AE}, d_{BE}) = d_{BE} = 7$.

Abbiamo cosi' ottenuto una nuova matrice di distanze $D^{(2)}$:

$D^{(2)} =$

	AB	C	D	E
AB	0	3	6	7
C	3	0	4	6
D	6	4	0	2
E	7	6	2	0

Il piu' piccolo elemento di $D^{(2)}$ vale 2, e rappresenta la distanza tra D ed E. Quindi le unità D ed E vengono aggregate in un cluster (DE).

Si puo' poi calcolare una nuova matrice di distanze $D^{(3)}$ e procedere nell'aggregazione. Avremo:

$$D^{(3)} = \begin{array}{c|ccc} & AB & C & DE \\ AB & 0 & 3 & 6 \\ C & 3 & 0 & 4 \\ DE & 6 & 4 & 0 \end{array}$$

Al passo successivo l'unità C viene aggregata al cluster (AB). Mentre nell'ultimo passo i due cluster (ABC) e (DE) vengono fusi in un unico gruppo.

Il metodo del **legame completo** si basa su un criterio di **distanza massima**.Ovvero, supponendo di avere 4 unità : A,B,C,D, e di aver definito un coefficiente di dissimilarità o una misura di distanza tra le unità (d_{AB}, d_{AC}, ……,d_{CD}); supponendo che le unità A e B vengano fuse in un solo cluster, la distanza tra il cluster (AB) e l'unità C e' definita come:

$$d_{(A,B)C} = \max(d_{AC}, d_{BC})$$

mentre la distanza tra il cluster (AB) ed il cluster (CD) viene definita come:

$$d_{(AB)(CD)} = \max(d_{AC}, d_{AD}\ d_{BC}, d_{BD}).$$

Nel metodo del **legame medio** la distanza tra cluster e' definita come media aritmetica delle distanze (o dissimilarità) tra tutte le possibili coppie di elementi appartenenti l'uno ad un cluster, l'altro ad un altro.
Dati 2 cluster A e B, contenenti, rispettivamente, n_A ed n_B unità , indichiamo con l'indice *i* il generico elemento del cluster A, e con l'indice *h* il generico elemento del cluster B, e con $d_{i,h}$ la loro distanza . La distanza tra A e B e' definita come:

$$d_{A,B} = 1/n_A n_B(\Sigma_i \Sigma_h d_{i,h})$$

Per esempio, supponiamo di avere le stesse 5 unità dell'esempio precedente, la cui matrice delle distanze e':

$D^{(1)}$ =

	A	B	C	D	E
A	0	1	5	6	8
B	1	0	3	8	7
C	5	3	0	4	6
D	6	8	4	0	2
E	8	7	6	2	0

Nel primo passo vengono fuse le unità A e B, essendo le piu' "vicine" (con $d_{AB} = 1$).
A questo punto bisogna calcolare le distanze tra questo cluster (AB) e le altre unità.
$d_{(AB),C} = (1/2)x1(d_{A,C} + d_{B,C}) = 0.5(5+ 3) = 4$
$d_{(AB),D} = (1/2)x1 (d_{A,D} + d_{B,D}) = 0.5(6 + 8) = 7$
$d_{(AB),E} = (1/2)x1 (d_{A,E} + d_{B,E}) = 0.5(8 + 7) = 7.5$

La nuova matrice delle distanze sarà:

$D^{(2)} =$

	AB	C	D	E
AB	0	4	7	7.5
C	4	0	4	6
D	7	4	0	2
E	7.5	6	2	0

Al passo successivo vengono aggregate in un unico cluster le unità D ed E.
Le nuove distanze sono:

$d_{(AB),\,(DE)} = (1/2)x(1/2)(d_{A,D} + d_{A,E} + d_{B,D} + d_{B,E} =$
$= 0.25(6 + 8 + 8 + 7) = 7.25$

$d_{C,(DE)} = 1x(1/2)\,(d_{C,D} + d_{C,E}) = 0.5(4 + 6) = 5$

La nuova matrice delle distanze e':

$D^{(3)} =$

	AB	C	DE
AB	0	4	7.25
C	4	0	5
DE	7.25	5	0

Infine l'unità C viene aggregata al cluster (AB), ottenendo la stessa gerarchia ottenuta con il metodo del legame semplice (sono pero' cambiate le distanze e quindi le altezze dei rami del dendrogramma).
Non sempre i diversi metodi producono le stesse gerarchie.

Il **metodo del centroide** si applica solo a variabili quantitative e lavora non tanto sulla matrice delle distanze quanto sui singoli vettori di osservazioni. (Nel senso che ad ogni passo ricalcola la matrice delle distanze partendo non dalle distanze precedenti ma dai baricentri di ciascun cluster).
Per ogni gruppo (anche composto da una sola unità) si calcola il baricentro (o *individuo medio*, cioe' un elemento che come modalità delle diverse variabili, presenta le modalità medie del gruppo). La distanza tra un'unità e un gruppo o tra due gruppi e' calcolata come distanza tra i baricentri.
Esempio: supponiamo di avere 5 unità statistiche e due variabili; che assumono le seguenti modalità:

	x_1	x_2
A	1	1
B	1	2
C	6	3
D	8	2
E	8	0

La matrice dei quadrati delle distanze euclidee e' la seguente:

$D^{(1)} =$

	A	B	C	D	E
A	0	1	29	50	50
B		0	26	49	53
C			0	5	13
D				0	4
E					0

Infatti, per esempio, $d^2_{A,E} = (1-8)^2 + (1-0)^2 = 50$.

Le unità statistiche A e B hanno minore distanza e quindi si raggruppano in un primo cluster (AB).
Il baricentro del cluster (AB) avrà coordinate (1; 1.5),
il che e' come assumere che le modalità siano:

	x_1	x_2
AB	1	1.5
C	6	3
D	8	2
E	8	0

e la matrice dei quadrati delle distanze euclidee sarà:

$D^{(2)} =$

	AB	C	D	E
AB	0	27.25	49.25	51.25
C		0	5	13
D			0	4
E				0

In cui, per esempio, $d_{(AB)C}$ e' ottenuta come:
$d^2_{(AB)C} = (1-6)^2 + (1.5- 3)^2 = 25 + 2.25 = 27.25$

Si procede poi all'aggregazione delle unità D ed E, il cui baricentro e' dato da: (8; 1). Si ha cosi':

	x_1	x_2
AB	1	1.5
C	6	3
DE	8	1

E quindi:

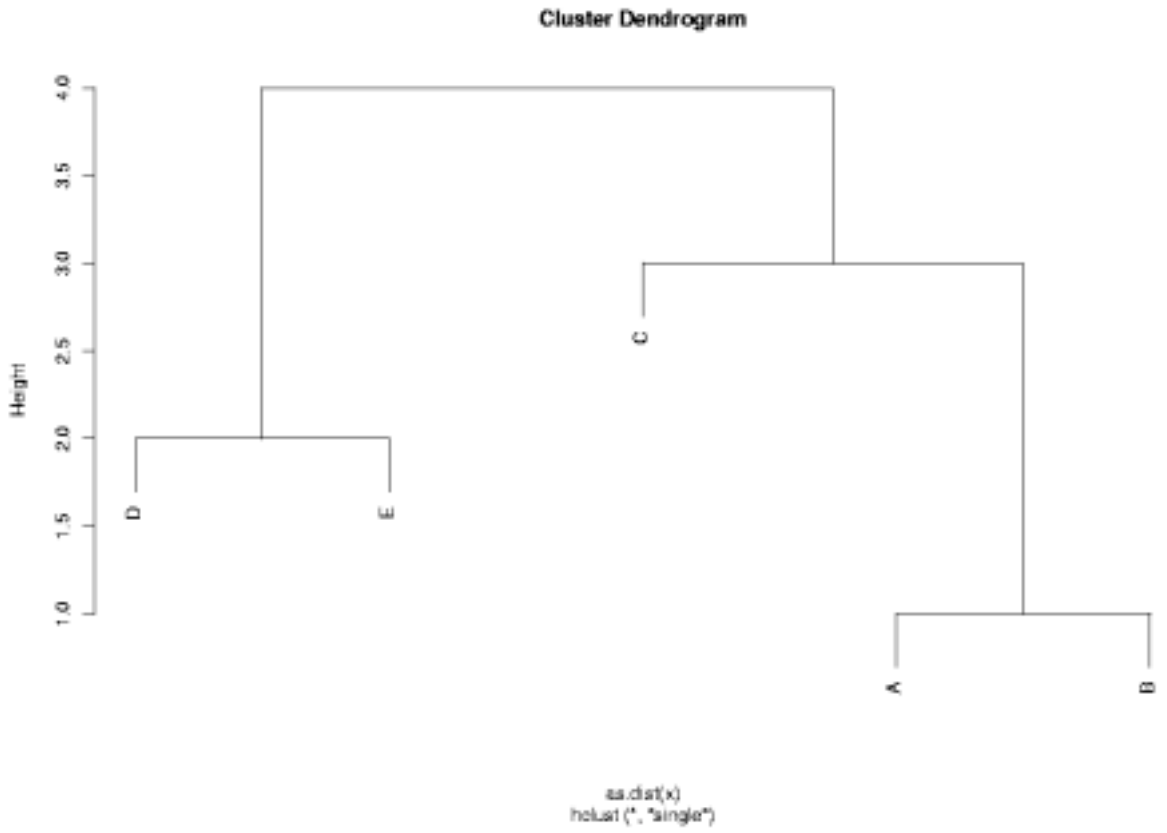
$D^{(3)} =$

	AB	C	D
AB	0	27.25	49.25
C		0	8
DE			0

Nell’ultimo passo l’unità C verrà aggregata al cluster (AB).

3.1 Il dendrogramma

Il dendrogramma e’ una rappresentazione grafica che visualizza secondo ordinate crescenti il livello di aggregazione delle unità o cluster.
Il dendrogramma dell’esempio precedente riguardante il metodo del legame singolo e’ il seguente:



In sostanza visualizza l’intero processo di aggregazione ossia una gerarchia di partizioni. Una singola partizione si ottiene “tagliando” il dendrogramma ad un dato livello dell’**indice di distanza della gerarchia**.

La scelta di quanti gruppi finali ottenere si traduce nel problema: **a quale livello tagliare l'albero?**
Dato che si ha interesse ad avere il **minor** numero di gruppi con **massima** omogeneità, si cerca di tagliare "alle radici" (cioe' in basso) dell'insieme dei "rami" piu' lunghi (cioe' le verticali piu' lunghe). (Nel caso precedente potremmo forse prenderci 3 cluster: (AB), C, (DE), oppure 2: (AB), (CDE)).

Definizione: partizione.
Una partizione P(E) su un insieme E si definisce come l'insieme delle classi di E tale che**:**

1. Due elementi A_i e A_h di P(E) sono o disgiunti (cioe' $A_i \cap A_h = \varnothing$) oppure coincidenti (cioe' $A_i \cup A_h = A_i = A_h$);
2. L'unione di tutte le parti esaurisce E (cioe' $A_1 \cup A_2 \cup \ldots \cup A_k$).

Definizione: gerarchia
Una sequenza (discendente) di partizioni $(P_1,\ldots..,P_s)$ di un insieme E, forma una gerarchia se e solo se per ogni P_q e P_s, con s >q, ogni elemento A_i di P_q e' contenuto o coincide con un elemento A_h di P_s. (Il che in sostanza significa che gli elementi di P_q vengono aggregati tra loro per arrivare alla partizione successiva.)

3.2Il metodo di Ward

Anche questo metodo puo' essere utilizzato come algoritmo gerarchico aggregativo.
Tale metodo e' diretto alla minimizzazione della varianza all'interno dei gruppi.(Pertanto **puo' essere utilizzato solo per variabili quantitative**). Ad ogni passo questo algoritmo tende ad ottimizzare la partizione ottenuta tramite l'aggregazione di due elementi.
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). In maniera analoga si scompone la matrice di varianze e covarianze S.
In simboli:

$$S = S_W + S_B$$

Dove S e' la matrice di varianze e covarianze totali;
S_W e' la matrice delle varianze e covarianze "interne";
S_B e' la matrice delle varianze e covarianze "esterne".

Parentesi: la scomposizione della devianza.

Il caso uni-variato.

Supponiamo che su una variabile x, si abbiano n osservazioni . La varianza è data da:

$$\Sigma_i(x_i - M_1(x))^2/n \qquad (1)$$

e la devianza e' semplicemente il suo numeratore.

Supponiamo ora che la variabile x rappresenti l'altezza di n individui, e che in realtà essi siano distinti in maschi e femmine. Indichiamo con x_{iF} le osservazioni che si riferiscono alle femmine e con x_{iM} le osservazioni che si riferiscono ai maschi.

Possiamo anche scrivere che la devianza totale e' data dalla somma su tutti i maschi degli scostamenti al quadrato tra le x_{iM} e la $M_1(x)$, e la somma su tutte le femmine di scostamenti analoghi.
Effettuando alcuni passaggi algebrici, definendo con n_M la numerosità dei maschi e con n_F la numerosità delle femmina (con $n_M + n_F = n$), e introducendo un indice j che vale M oppure F, si può dimostrare che il numeratore della (1) equivale anche alla somma di due quantità:

$$\Sigma_j\Sigma_i(x_{ij} - M_1(x_j))^2 + \Sigma_j\Sigma_i (M_1(x_j) - M_1(x))^2 \qquad (2)$$

La prima quantità rappresenta la **devianza interna**. Se sviluppiamo rispetto alla somma in j, la devianza interna risulta a sua volta una somma di due devianze (la prima e' una devianza interna alle femmine, la seconda e' una devianza interna al gruppo di maschi):

$$\text{Dev. Interna} = \Sigma_i(x_{iM} - M_1(x_M))^2 + \Sigma_i(x_{iF} - M_1(x_F))^2$$

Il secondo addendo della (2) rappresenta invece la **devianza esterna** (la quale misura la variabilità tra le medie dei due gruppi).
Ricordando che l'indice j= M, F, e che l'indice i = 1,….,n_M oppure i= 1,….,n_F, la devianza esterna puo' essere riscritta come somma di due quantità:

$$\text{Dev. Est.}=(M_1(x_M) - M_1(x))^2 n_M + (M_1(x_F) - M_1(x))^2 n_F$$

La (2) rappresenta la cosi' detta "scomposizione della devianza". Vale per un numero qualsiasi di gruppi, cioe' puo' essere immediatamente estesa al caso in cui sia j=1,…,g (di conseguenza sarà i= 1,….,n_j).
Volendo scomporre la varianza anziché la devianza, basterà dividere per n sia la devianza interna che la devianza esterna.

Osservazioni:

- Il raggruppamento in classi di alcune o di tutte le unità non modifica la loro varianza complessiva. Si modificano solo i due addendi (varianza interna ed esterna), ma la loro somma non varia.
- All’aumentare del numero di unità contenute in un cluster, aumenta la varianza interna al cluster.Infatti in un cluster di un solo elemento la varianza interna nulla (perché l’unico elemento rappresenterà anche il baricentro del cluster). Un cluster di due elementi non coincidenti, avrà una varianza positiva. E cosi’ via: aggiungendo elementi ad un cluster, la varianza interna al cluster aumenta.
- Al diminuire del numero di cluster diminuisce la varianza esterna, in quanto risulterà pari alla somma di un minor numero di termini.
- Nel caso limite di un solo cluster contenente tutte le unità, la varianza esterna e’ nulla e la varianza interna coincide con la varianza totale.

Qualora si abbiano piu' variabili (p), una scomposizione di questo tipo vale per l'intera matrice S di varianze e covarianze. Sarà quindi:

$$S = S_W + S_B$$

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.

4.**Metodi gerarchici scissori**

Tali metodi definiscono partizioni sempre piu' fini dell'insieme iniziale; si suddivide progressivamente l'insieme E in un numero sempre crescente di sottoinsiemi, fino ad ottenere tutti i suoi elementi distinti. Tali metodi si basano sulla partizione di un insieme in due sottoinsiemi, e sulla suddivisione delle classi precedentemente ottenute, sempre e soltanto in ulteriori bipartizioni.

Anche questi metodi di solito partono dalla scomposizione della devianza, in particolare della matrice di devianze e codevianze (di solito indicata con T).

Sia T = W +B.

Il metodo di **Edwards e Cavalli-Sforza** sceglie come funzione obiettivo (da minimizzare) la traccia della matrice W. Ad ogni step si effettua la bipartizione che minimizza la traccia della matrice W, cioe' la somma delle devianze interne).

Il metodi di **Friedman e Rubin** ha come obiettivo la minimizzazione del determinante di W. Ma anche – in una variante – la traccia della matrice BW^{-1} .

5 Problemi aperti

- scelta del numero di cluster
- scelta delle variabili
- standardizzazione o eliminazione dell'u.d.m.

6 Metodi non gerarchici

Caratteristiche:

- Sono di solito algoritmi aggregativi e producono una sola partizione.
- Ad ogni passo dell'algoritmo rimettono in discussione la partizione ottenuta. Le classi ottenute ad ogni iterazione intermedia vengono infatti cancellate e il processo di aggregazione ricomincia, a partire dai nuovi centri.
- L'inizializzazione del processo di classificazione e' necessariamente data da una qualche scelta di un insieme di g centri iniziali.

Gli algoritmi gerarchici risultano spesso eccessivamente onerosi in termini di calcoli che richiedono. Un modo per ridurre la quantità di calcoli, e' di scegliere a priori il numero di cluster. (In tal modo l'algoritmo non sarà piu' gerarchico, e genererà un'unica partizione delle n unità in g clusters.)

Mentre nel caso dei metodi gerarchici l'algoritmo cerca, ad ogni passo, la migliore scissione o aggregazione tra cluster, nel caso dei metodi non gerarchici l'algoritmo partiziona le unità in un numero predefinito di gruppi basandosi sulla ottimizzazione di un qualche criterio (predefinito). L'inizializzazione dell'algoritmo avviene indicando g centri di partenza intorno a cui aggregare le unità. A differenza dei metodi gerarchici, **l'assegnazione di un oggetto ad un cluster non e' irrevocabile**. Ovvero le unità vengono riassegnate ad un diverso cluster se l'allocazione iniziale risulta inappropriata.

Anche in questo caso esistono diversi algoritmi.
Differiscono tra loro nei seguenti aspetti:

1. come sono inizializzati i centri di partenza;
2. come gli elementi vengono assegnati ai diversi centri;
3. come alcune o tutte le unità vengono eventualmente riassegnate ad un diverso gruppo.

Di solito, scelta una partizione iniziale, si cerca di migliorarla in funzione del criterio di minimizzazione della varianza interna.

6.1Descrizione generale di un algoritmo non gerarchico

Inizializzazione: si individuano g centri provvisori, i quali inducono una prima partizione provvisoria. Tale classificazione avviene sulla base della minima distanza di un individuo da uno di questi centri. Si calcola, come nel metodo del centroide, il baricentro di ogni gruppo, cioe' *l'individuo medio*. Per ogni gruppo, si calcola la varianza interna: la loro somma viene indicata con W_0.
Primo passo: si assumono i baricentri appena calcolati come nuovi centri provvisori. Si ripete il procedimento di allocazione delle unità ai centri sulla base della minima distanza. Si calcolano i nuovi baricentri delle classi (*individui medi*). E di nuovo, la somma delle varianze interne ad ogni classe: W_1.
Passi successivi: ad ogni iterazione successiva si cancella la partizione ottenuta in precedenza e si reitera il processo di aggregazione, assumendo come nuovi centri provvisori i baricentri del passo precedente.

Stop: se non vi sono state riallocazioni.
Oppure: se $W_{t-1} - W_t$ e' inferiore a soglia prefissata.

Si puo' dimostrare che ad ogni iterazione la varianza interna alle classi non puo' che diminuire.

Si basano su un algoritmo del tipo appena descritto sia il metodo delle **aggregazioni dinamiche**, che il metodo **k-means**.
Nel metodo delle aggregazioni dinamiche si scelgono g centri provvisori tramite estrazione casuale dalle n unità. Si decide la regola di stop fissando una soglia alla differenza tra $W_{t-1} - W_t$.
Nel metodo k-means si assumono come centri provvisori i primi k individui. Si allocano via via le n-k unità e ad ogni assegnazione si ricalcola subito il centroide del gruppo che si e' modificato. In tal modo si accelera il miglioramento della classificazione. Si calcola la varianza interna e si passa allo step successivo prendendo i baricentri dei gruppi appena ottenuti. La regola di stop si basa sulla differenza tra $W_{t-1} - W_t$.

Problema: l'algoritmo potrebbe convergere ad un **ottimo locale (e non globale)**. Il che significa che se partissimo da un diverso insieme di centri provvisori, potremmo ottenere una partizione differente.

6.2Variante: i gruppi stabili.

Per ovviare il problema di convergenza ad un ottimo locale, si puo' andare alla ricerca dei gruppi stabili. In sostanza si ripete piu' volte un intero processo di classificazione, partendo da pochi nuclei iniziali. I diversi processi di classificazione saranno diversi proprio nella scelta dei nuclei iniziali. Si incrociano poi i risultati delle diverse analisi, per individuare i gruppi stabili, ovvero quei cluster che risultano dall'intersezione delle diverse partizioni.

Volendo, questi gruppi stabili possono essere presi come nuclei iniziali di un'ulteriore strategia di classificazione.