Clustering iterativo
Approccio descrittivo
Si dispone di "*m*" informazioni di tipo metrico su ciascuna di "*n*" entità di un collettivo:

$$x = \{x_1, x_2, ..., x_n\}$$

In cluster analysis ha come obiettivo la definizione di una partizione di *X* in un numero fissato "k" di sottoinsiemi, o gruppi, o cluster C_i, tali che

$$C_1 \cup C_2 \cup ... \cup C_k = X \qquad C = \{c_1, c_2, ..., c_k\} \in P(n,k)$$

$$C_i \cap C_j = \phi$$

$$i \neq j$$

Inoltre, se n_i indica il numero di entità appartenenti al cluster C_i, allora si deve avere

$$1 < n_i < n - k + 1 \qquad \sum_{i=1}^{k} n_i = n$$

Esempio:

$$x = \{x_1 x_2, ..., x_s\}$$

$$\begin{array}{ll} C_1 = \{x_2, x_3, x_7\} & n_1 = 3 \\ C_2 = \{x_1, x_4\} & n_2 = 2 \\ C_3 = \{x_5, x_6 x_9\} & n_3 = 3 \\ C_4 = \{x_s\} & n_4 = 1 \end{array} \qquad n = 9$$

Il numero di partizioni possibili è dato da

$$\frac{1}{k!}\sum_{j-1}^{k}(-1)^{k-j}\binom{k}{j}j^m$$

Che esprime il numero di elementi in *P(n,k)* cioè l'insieme delle possibili partizioni in "*k*" gruppi delle "*n*" entità.
Per *n = 9* e g = 4 abbiamo:

$$\frac{1}{24}\{-4 + 6 \cdot 2^9 - 4 \cdot 3^9 + 4^9\} = 7770$$

che è già un numero elevato. Per *n=100* e *g=5* il numero di partizioni eccede 10^{68} per cui se disponesse di un miliardo di computer ciascuno con un miliardo di processori in grado ciascuno di esaminarne un miliardo di miliardi in un nanosecondo, per eseaminarle tutte sarebbero necessari

diecimila miliardi di anni (pare che il nostro universo esista "solo" da 13 miliardi di anni. È per questo che occorre fare ricorso a dei metodi approssimati ed iterativi in modo analogo con ciò che si fa nel risolvere i sistemi di equazioni non lineari di tipo continuo.

Decomposizione della matrice di dispersione totale
Per una generica entità x_i ed un punto indeterminato y si ha l'ovvia identità:

$$(x_i - y) = (x_i - \mu_j) + (\mu_j - y)$$

Dove μ_j è il vettore delle medie o centroide del cluster j-esimo.

$$\mu_j = \frac{1}{n_j} \sum_{x_i \in C_j} x_i$$

Costruiamo ora il prodotto esterno

$$\begin{aligned}(x_i - y)(x_i - y)' &= [(x_i - \mu_j) + (\mu_j - y)][(x_i - \mu_j) + (\mu_j - y)] \\ &= (x_i - \mu_j)(x_i - \mu_j) + (\mu_j - y)(\mu_j - y) + 2(x_i - \mu_j)(\mu_j - y)\end{aligned}$$

Ne consegue che la somma su tutto il cluster j-esimo

$$\sum_{x_i \in C_j} (x_i - y)(x_i - y)'$$

Diviene:

$$\sum_{x_i \in C_j} (x_i - \mu_j)(x_i - \mu_j) + n_j (\mu_j - y)(\mu_j - y) + 2\left(\sum_{x_i \in C_j} (x_i - \mu_j) \right)(n_i - y)'$$

Per cui, tenendo conto della proprietà della media aritmetica, si ha

$$\sum_{x_i \in C_j} (x_i - y)(x_i - y)' = \sum_{x_i \in C_j} (x_i - \mu_j)(x_i - \mu_j) + n_j (\mu_i - y)(\mu_i - y)'$$

Se ora poniamo y=μ dove

$$\mu = \sum_{i=1}^{n} x_i / n$$

E scriviamo

$$W_j \sum_{y_i \in C_j} (x_i - \mu_j)(y_i - \mu_j); \qquad B_j = n_j (\mu_j - \mu)(\mu_j - \mu)$$

Otteniamo l'identità

$$T_j = W_j + B_j$$

Estendiamo ora la bipartizione a tutti i "*y*" gruppi. Si arriva alla decomposizione della matrice di dispersione totale:

$$\begin{aligned} T &= \sum_{y=1}^{g} T_j = \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^t = W + B \\ &= \sum_{j=1}^{g} W_j + \sum_{j=1}^{g} B_5 \\ &= \sum_{j=1}^{g} \sum_{x_i \in C_j} (x_i - \mu_j)(x_i - \mu_j) + \sum_{j=1}^{g} n_j (\mu_j - \mu)(\mu_j - \mu) \end{aligned}$$

N.B. T, W, e B sono tutte simmetriche

$$\begin{aligned} T &= W + B \\ &= \sum_{j=1}^{g} W_j + \sum_{j=1}^{g} B_5 \\ &= \sum_{j=1}^{g} \sum_{x_i \in C_j} (x_i - \mu_j)(x_i - \mu_j)^t + \sum_{j=1}^{g} n_j (\mu_j - \mu)(\mu_j - \mu)^t \end{aligned}$$

La matrice W esprime la dispersione all'interno di ogni cluster ed è detta componente "within". La matrice "B" esprime la dispersione tra i cluster rappresentati dai loro centroidi ed è detta componente "between".

La suddivisione della varianza ne consente una nuova interpretazione: se i gruppi avessero la stessa media non ci sarebbe variabilità "between" e la variabilità complessiva deriverebbe solo da diversificazioni interne ai gruppi; d'altra parte, se i gruppi presentassero sempre le stesse modalità sparirebbe la variabilità "within" e conterebbero solo le differenze tra i centroidi dei gruppi. Possiamo avere perciò variabilità perché i gruppi differiscono al loro interno o perché differiscono tra di loro o per entrambe le ragioni.

Da notare che per un ogni data set, la matrice di dispersione totale T è fissa e invariabile per cui ogni modifica in *W* deve trovare una modifica in *B* di segno opposto, ma di entità equivalente.

Esempio

$$C_1 = \left\{\begin{bmatrix}3\\-3\end{bmatrix}\begin{bmatrix}4\\-2\end{bmatrix}\begin{bmatrix}2\\-1\end{bmatrix}\right\}; \mu_1 = \begin{bmatrix}3\\-2\end{bmatrix};$$

$$C_2 = \left\{\begin{bmatrix}4\\2\end{bmatrix}; \begin{bmatrix}2\\4\end{bmatrix}\right\}; \mu_2 = \begin{bmatrix}3\\3\end{bmatrix}$$

$$\mu = \begin{bmatrix}3\\0\end{bmatrix}$$

$$(x_1 - \mu) = \begin{bmatrix}3 & -3\\-3 & 0\end{bmatrix} (x_2 - \mu) = \begin{bmatrix}4 & -3\\-2 & -0\end{bmatrix}; (x_3 - \mu) = \begin{bmatrix}2 & -3\\-1 & -0\end{bmatrix}; (x_4 - \mu) = \begin{bmatrix}4 & -3\\2 & -0\end{bmatrix}; x_5 = \begin{bmatrix}2 & -3\\4 & -0\end{bmatrix}$$

$$T = \begin{bmatrix}0\\-3\end{bmatrix}\begin{bmatrix}0 & -3\end{bmatrix} + \begin{bmatrix}1\\-2\end{bmatrix}\begin{bmatrix}1 & -2\end{bmatrix} + \begin{bmatrix}-1\\-1\end{bmatrix}\begin{bmatrix}1 & -1\end{bmatrix} + \begin{bmatrix}1\\2\end{bmatrix}\begin{bmatrix}1 & 2\end{bmatrix} + \begin{bmatrix}-1\\4\end{bmatrix}\begin{bmatrix}-1 & 4\end{bmatrix}$$

$$= \begin{bmatrix}0 & 0\\0 & 9\end{bmatrix} + \begin{bmatrix}1 & -2\\-2 & 4\end{bmatrix} + \begin{bmatrix}1 & 1\\1 & 1\end{bmatrix} + \begin{bmatrix}1 & 2\\2 & 4\end{bmatrix} + \begin{bmatrix}1 & -4\\-4 & 16\end{bmatrix} = \begin{bmatrix}4 & -3\\-3 & 34\end{bmatrix}$$

$$(x_1 - \mu) = \begin{bmatrix}3 & -3\\-3 & +2\end{bmatrix}; (x_2 - \mu_1) = \begin{bmatrix}4 & -3\\-2 & +2\end{bmatrix}; (x_3 - \mu_1) = \begin{bmatrix}2 & -3\\-1 & +2\end{bmatrix}; (x_4 - \mu_2) = \begin{bmatrix}4 & -3\\2 & -3\end{bmatrix}; (x_5 - \mu_2) = \begin{bmatrix}2 & -3\\4 & -3\end{bmatrix}$$

$$W_1 = \begin{bmatrix}0\\-1\end{bmatrix}\begin{bmatrix}0 & -1\end{bmatrix} + \begin{bmatrix}1\\0\end{bmatrix}\begin{bmatrix}1 & 0\end{bmatrix} + \begin{bmatrix}-1\\1\end{bmatrix}\begin{bmatrix}-1 & 1\end{bmatrix} = \begin{bmatrix}0 & 0\\0 & 1\end{bmatrix} + \begin{bmatrix}1 & 0\\0 & 0\end{bmatrix} + \begin{bmatrix}1 & -1\\-1 & 1\end{bmatrix} = \begin{bmatrix}2 & -1\\-1 & 2\end{bmatrix}$$

$$W_2 = \begin{bmatrix}1\\-1\end{bmatrix}\begin{bmatrix}1 & -1\end{bmatrix} + \begin{bmatrix}-1\\1\end{bmatrix}\begin{bmatrix}-1 & 1\end{bmatrix} = \begin{bmatrix}1 & -1\\-1 & 1\end{bmatrix} + \begin{bmatrix}1 & -1\\-1 & 1\end{bmatrix} = \begin{bmatrix}2 & -2\\-2 & 2\end{bmatrix}$$

$$B_1 = 3\begin{bmatrix}3 & -3\\-2 & -0\end{bmatrix}\begin{bmatrix}0 & -2\end{bmatrix} = 3\begin{bmatrix}0 & 0\\0 & 4\end{bmatrix} = \begin{bmatrix}0 & 0\\0 & 12\end{bmatrix}$$

$$B_2 = 2\begin{bmatrix}3 & -3\\3 & -0\end{bmatrix} \cdot \begin{bmatrix}0 & 3\end{bmatrix} = 2\begin{bmatrix}0 & 0\\0 & 9\end{bmatrix} = \begin{bmatrix}0 & 0\\0 & 18\end{bmatrix}$$

$$W = \begin{bmatrix}2 & -1\\-1 & 2\end{bmatrix} + \begin{bmatrix}2 & -2\\-2 & 2\end{bmatrix} = \begin{bmatrix}4 & -3\\-3 & 4\end{bmatrix}$$

$$B = \begin{bmatrix}0 & 0\\0 & 4\end{bmatrix} + \begin{bmatrix}0 & 0\\0 & 18\end{bmatrix} = \begin{bmatrix}0 & 0\\0 & 22\end{bmatrix}$$

$$W + B = \begin{bmatrix}4 & -3\\-3 & 4\end{bmatrix} + \begin{bmatrix}0 & 0\\0 & 30\end{bmatrix} = \begin{bmatrix}4 & -3\\-3 & 34\end{bmatrix} = T$$

Compattezza euclidea del cluster e ottimalità rispetto alla devianza. Minimo di Tr(W)

La quantità

$$e_j^2 = \sum_{x_i \in C_j} \frac{\left\| x_i - \mu_j \right\|^2}{n_j} \quad j = 1,2,\ldots,k$$

È pari alla media delle distanze euclidee degli elementi appartenenti al cluster j-esimo. Con essa si può avere una idea di quanto ravvicinate siano le unità intorno al loro centroide.
Se le compattezze di cluster vengono sommate su tutte le "*n*" entità otteniamo

$$D(C) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \frac{\left\| x_i - \mu_j \right\|^2}{n} = \sum_{j=1}^{k} \left(\frac{n_j}{n} \right) e_j^2$$

Che può misurare la qualità di una data partizione.
Una data partizione $C_1, C_2, \ldots, C_n = C_0$ e $P(n,k)$ (questo simbolo indica l'insieme di tutte le partizioni di entità in "k" gruppi) è ottima rispetto alla devianza se

$$D(C_0) = \min_{C \in P(n,k)} D(C) = \sum_{j=1}^{k} \left(\frac{n_j}{n} \right) e_j^2 \qquad n_j \geq 1, \sum_{j=1}^{k} n_j = n$$

Il criterio può anche essere formulato in modo più comprensbibile. Dalla dispersione totale:

$$\left(\frac{1}{n} \right) Tr(T) = \frac{1}{n} Tr\left(\sum_{j=1}^{k} T_j \right) = \frac{1}{n} \sum_{j=1}^{k} Tr(T_j)$$

Si arriva alla suddetta scomposizione dato che la traccia di una somma è pari alla somma delle tracce. D'altra parte $T_j = W_j + B_j$ perciò

$$\begin{aligned} D(C) &= \frac{1}{n} \left\{ \sum_{j=1}^{g} Tr(w_j) \right\} + \frac{1}{n} \sum_{j=1}^{g} Tr(B_j) \\ &= \frac{1}{n} \sum_{j=1}^{g} n_j e_j^2 + \frac{1}{n} \sum_{j=1}^{n} n_j \left\| \mu_j - \mu \right\|^2 \\ &= \sum_{j=1}^{g} \left(\frac{n_j}{n} \right) e_j^2 + \sum_{j=1}^{g} \left(\frac{n_j}{n} \right) \left\| \mu_j - \mu \right\|^2 \end{aligned}$$

Poiché la dispersione totale di un data set (blocco di dati) è fissa possiamo determinare la partizione ottima secondo due modi diversi, ma con soluzione equivalente:

$$\min_{C \in P(n,k)} \sum_{j=1}^{k} \left(\frac{n_j}{n} \right) e_j^2; \qquad \max_{C \in P(n,k)} \sum_{j=1}^{k} \left(\frac{n_j}{n} \right) \left\| n_j - \mu \right\|^2$$

Ne consegue che muovere alla ricerca della partizione che minimizza la devianza interna dei cluster, è la stessa cosa di cercare la partizione che massimizza la distanza tra i centroidi dei clusters. Si rispetta perciò il principio di costruire una partizione in cui le entità incluse in un dato cluster sono omogenee tra di loro e, nello stesso tempo, diverse da quelle in altri clusters.

Per una più agevole spiegazione del criterio della minimizzazione della traccia consideriamo la relazione:

$$\sum_{x_i \in C_j} (x_i - x_k)(x_i - x_k)' = \sum_{x_i \in C_j} (x_i - \mu_j)(x_i - \mu_j)' + n_j (\mu_j - x_k)(\mu_j - x_k)'$$

Se ora sommiamo rispetto a tutte le si ha

$$x_k \in C_j$$

$$\begin{aligned} \sum_{x_k \in C_j} \sum_{x_i \in C_j} (x_i - x_k)(x_i - x_k)' &= \sum_{x_k \in C_j} \sum_{x_i \in C_j} (x_i - \mu_j)(x_i - \mu_j)' + n_j \sum_{x_k \in C_j} (\mu_j - x_k)(\mu_j - x_k)' \\ &= 2n_j \sum_{x_k \in C_j} (\mu_j - x_k)(\mu_j - x_k)' \\ &= 2n_j W_j \end{aligned}$$

ovvero

$$W_j = \frac{1}{2n_j} \sum_{x_k \in C_j} \sum_{x_i \in C_j} (x_i - x_k)(x_i - x_k)'$$

per cui la dispersione all'interno di un cluster è pari alla distanza media tra due entità qualsiasi all'interno di ogni cluster. Da notare che secondo questa formulazione i centroidi dei cluster non compaiono in modo esplicito.

Minimizzazione della traccia di *W*

Una data partizione

$$C = \{C_1, \ldots, C_k\}$$

È la partizione a distanza minima rispetto alla *Tr(W)* se

$$\left\| x_i - \mu_j \right\| = \min_{g=1,2,\ldots,k} \left\{ \left\| x_i - \mu_g \right\| \right\} \quad se \quad x_i \in C_j$$

Cioè ogni entità dista dal proprio centroide meno di quanto non disti dal centroide di ogni altro cluster.

Data la relazione che sussiste tra la distanza euclidea e la partizione ottima rispetto a *Tr(W)* è anche la partizione a distanza minima. Il contrario non è sempre vero.

Algoritmo di calcolo

Sia data una partizione iniziale

$$C^{(0)} = \left\{C_1^{(0)}, C_2^{(0)}, \ldots, C_k^{(0)}\right\} \qquad n_j^{(0)} \geq 1 \qquad \sum_{j=1}^{k} n_j^{(0)} = n.$$

Passo_1: si calcolano i vettori delle medie

$$\mu_j^{(0)} = \frac{1}{n_j^{(0)}} \sum_{X_i \in C_j^{(k)}} X_i$$

Passo_2: si determina l'attuale partizione a distanza minima

$$C^{q+1} = \left\{C_1^{(q+1)}, C_2^{(q+1)}, \ldots, C_k^{(q+1)}\right\}$$

Dove

$$x_i \in C_j^{(q+1)} \quad se \quad \left\|x_i - \mu_j^{(q)}\right\| = \min_{g=1,2,\ldots,k} \left\|x_i - \mu_g^{(q)}\right\|$$

Si ritorna al passo 2.

Da notare che il ritorno al passo due può avvenire in varie fasi. Ad esempio dopo ogni confronto con il ricalcolo dei centroidi coinvolti oppure dopo la escursione di tutte le "*n*" entità. L'aggiornamento complessivo (cioè l'effettuazione del passo 1 dopo che per tutte le entità si sia verficato il passo 2 senza aggiornare i centroidi) è più rapido, implica minori calcoli, ma richiede più ripetizioni e rischia più spesso di non determinare la partizione ottima (Passo di Forgy).

Un algoritmo più efficace può essere ottenuto effettuando il ricalcolo delle medie ogni volta che si realizza una riassegnazione. L'effetto di un trasferimento è facilmente quantificato (Passo di McQueen).

$$x_i \in C_\mu$$

Supponiamo che

Debba essere assegnato a C_e cosicché C_μ perde una entità e C_e ne guadagna una:

$$C_\mu^{(y+1)} = \frac{C_\mu^{(y)}}{x_i}$$

e

$$C_e^{(y+1)} = C_e^{(y)} \cup x_i$$

$$\mu_\mu^{(y+1)} \frac{1}{n_\mu - 1}\left(n_\mu \mu_\mu^{(y)} - x_i\right)$$

$$\mu_e^{(y+1)} \frac{1}{n_e + 1}\left(n_e \mu_e^{(y)} - x_i\right)$$

L'effetto sulla traccia di W è il seguente:

$$W^{(y+1)} - W^{(y)} = \sum_{j=1}^{g} W^{(y+1)} - \sum_{j=1}^{g} W^{(y)} = W_e^{(y+1)} - W_e^{y} + W_\mu^{y}$$

$$Tr\left(W_e^{y+1}\right) = Tr\left(W_e^{(y)}\right) + \frac{n_e}{n_e + 1}\left\|x_i - \mu_e^{(y)}\right\|^2;$$

$$Tr\left(W_\mu^{y+1}\right) = W_\mu^{(y)} - \frac{n_\mu}{n_\mu - 1}\left\|x_i - \mu_\mu^{y}\right\|^2$$

E quindi

$$Tr\left[W^{(y+1)} - W^{(y)}\right] = \frac{n_e}{n_e + 1}\left\|x_i - \mu_e^{(y)}\right\|^2 - \frac{n_\mu}{n_\mu - 1}\left\|x_i - \mu_\mu^{(y)}\right\|^2$$

Ci sarà perciò un miglioramento del criterio se solo se

$$\frac{n_e}{n_e + 1}\left\|x_i - \mu_e^{(y)}\right\|^2 < \frac{n_\mu}{n_\mu - 1}\left\|x_i - \mu_\mu^{(y)}\right\|^2$$

Per

$q \to \infty$

Questo schema produce la partizione a distanza minima oltre che la partizione ottima rispetto a *Tr(W)*.

Algoritmo_TRWEXM

Passo_0:

E' data una partizione iniziale