CdL in Scienze Statistiche ed Economiche - Università degli Studi di Milano-Bicocca

Lezione : Metodo delle K-medie

Docente: Aldo Solari

1 Analisi dei gruppi

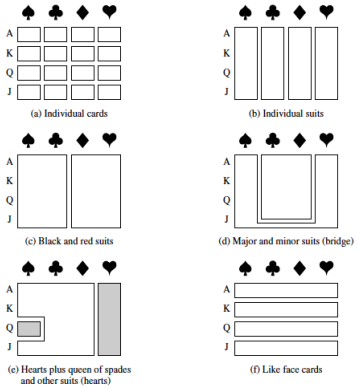
Perchè raggruppare? Suddividere le unità in gruppi è un modo naturale e, si può dire, imprescindibile, di ragionare per comprendere i fenomeni. Si ragiona per gruppi perchè è più facile dominare mentalmente pochi gruppi che tante unità. Uno stesso insieme di unità consente diversi raggruppamenti, nessuno è 'giusto', semmai può essere utile (o inutile o anche dannoso). Raggruppare utilmente: mettere insieme unità simili e separare unità dissimili, in altre parole creare gruppi:

- omogenei al loro interno (*internal cohesion*)
- disomogenei tra di loro (*external isolation*)

Nell'analisi di raggruppamento (o *cluster analysis*) la domanda cui si vuol rispondere è se esistono e quanti sono dei gruppi sensati (naturali) in cui suddividere le unità sulla base delle variabili osservate.

Se si ha una conoscenza approfondita del fenomeno in esame, si è in grado di distinguere tra 'buoni' raggruppamenti e 'cattivi' raggruppamenti. Perchè non semplicemente considerare tutti i possibili raggruppamenti (*partizioni* di n unità in K gruppi) e sceglierne il 'migliore'?

Example 1.1. $n = 16$ *carte con figura o asso per ciascun seme francese (picche, cuori, quadri e fiori)*



- *1 modo di formare 1 singolo gruppo* $K = 1$
- *32767 modi di formare 2 gruppi* $K = 2$

- *7141686 modi di formare 3 gruppi* $K = 3$
- *etc.*
- *1 modo di formare* n *gruppi* $K = n = 16$

per un totale di 10 480 142 147 partizioni possibili

1.1 Numero di partizioni possibili

Il numero di tutte le possibili partizioni di n unità in K gruppi è

$$S(n, K) = \frac{1}{K!} \sum_{k=0}^{K} (-1)^{K-k} \binom{K}{k} k^n$$

dove $S(n, K)$ è il numero di Stirling (di seconda specie), quindi si dovrebbero considerare un numero di possibilità pari all' n-esimo numero di Bell

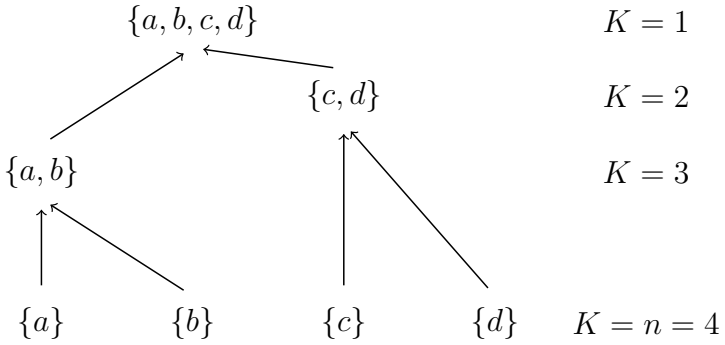
$$B_n = \sum_{K=1}^{n} S(n, K)$$

possibilità in tutto. Data la scarsa percorribilità dell'esplorazione di tutte le possibili partizioni, si procede usando dei metodi (algoritmi) che non esplorano l'intero spazio di tutte le possibili partizioni ma solo una parte di esse: non v'è perciò garanzia di ottenere la soluzione ottima in senso assoluto.

1.2 Metodi (algoritmi) gerarchici e non

Nei *metodi gerarchici* si individua una sequenza di partizioni nidificate: la partizione in $K + 1$ gruppi si ottiene dalla partizione in K gruppi facendo di due degli elementi di questa un elemento di quella (AGNES), o viceversa (DIANA)

- Algoritmo Agglomerativo (AGNES, AGGlomerative NESting)



- Algoritmo Scissorio (DIANA, DIvisive ANAlysis)

Nei *metodi non gerarchici*: il numero di gruppi K è deciso a priori

- Metodo delle K-medie

2 Metodo delle K-medie

Fissato a priori il numero dei gruppi K, come possiamo scegliere i gruppi

$$G_1, \ldots, G_K$$

in maniera 'ottimale'?
Si consideri la *distanza Euclidea al quadrato* tra due unità, i.e.

$$d^2(u_i, u_l) = \sum_{j=1}^{p} (x_{ij} - x_{lj})^2$$

La *distanza totale* risulta

$$T = \frac{1}{n} \sum_{i=1}^{n} \sum_{l=1}^{n} d^2(u_i, u_l)$$

Possiamo scomporre la distanza totale T in

$$T = W + B$$

dove

- B è la distanza tra i gruppi (*between*)
- W è la distanza entro i gruppi (*within*)

La distanza entro i gruppi si può esprimere come

$$W = \sum_{k=1}^{K} W(G_k)$$

dove

$$W(G_k) = \frac{1}{n_k} \sum_{i:u_i \in G_k} \sum_{l:u_l \in G_k} d^2(u_i, u_l)$$

è la distanza entro il k-simo gruppo, e n_k è la numerosità del gruppo G_k.

Vogliamo determinare i gruppi $G_1^*, \ldots, G_K^*$ tali che

$$W^* \leq W$$

ovvero risolvere il problema di minimo

$$\min_{G_1, \ldots, G_K} W$$

Si noti che determinare $G_1^*, \ldots, G_K^*$ che minimizza W comporta anche la massimizzazione di B poichè T è costante (non dipende dai $G_1, \ldots, G_K$).

Example 2.1. $u_1' = (0,0)$, $u_2' = (0,3)$, $u_3 = (4,3)$, *baricentro* $\underset{1\times 2}{\bar{x}'} = (4/3, 2)$

G_1, G_2	$\frac{1}{2}W$	$\frac{1}{2}B$	$\frac{1}{2}T$
{1}, {2,3}	*8*	*8.6*	*16.6*
{1,2}, {3}	*4.5*	*12.1*	*16.6*
{1,3}, {2}	*12.5*	*4.1*	*16.6*

Tuttavia, data la scarsa percorribilità dell'esplorazione di tutte le possibili partizioni, l'algoritmo delle K-medie non esplora l'intero spazio di tutte le possibili partizioni ma solo una parte di esse: non v'è perciò garanzia di ottenere la soluzione ottima in senso assoluto.
La distanza totale (in R, `totss`) può essere espressa come

$$\frac{1}{2}T = \sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - \bar{x}_j)^2 = \frac{1}{2}W + \frac{1}{2}B$$

dove $\bar{x}_j$ è il j-simo elemento del vettore delle medie $\bar{x}$.

Analogamente, la distanza entro il k-simo gruppo può essere espressa come

$$W(G_k) = 2\sum_{i:u_i\in G_k} d^2(u_i, \bar{x}_k) = 2\left[\sum_{i:u_i\in G_k}\sum_{j=1}^{p}(x_{ij} - \bar{x}_{kj})^2\right]$$

dove $\bar{x}_{kj}$ è il j-simo elemento del k-simo *centroide* (vettore delle medie del gruppo G_k).

$$\underset{p\times 1}{\bar{x}_k} = \begin{bmatrix} \bar{x}_{k1} \\ \cdots \\ \bar{x}_{kp} \end{bmatrix} = \begin{bmatrix} \frac{1}{n_k}\sum_{i:u_i\in G_k} x_{i1} \\ \cdots \\ \frac{1}{n_k}\sum_{i:u_i\in G_k} x_{ip} \end{bmatrix}$$

Per minimizzare la distanza entro i gruppi (in R, `tot.withinss`)

$$\frac{1}{2}W = \frac{1}{2}\sum_{k=1}^{K} W(C_k) = \sum_{k=1}^{K}\left[\sum_{i:u_i\in G_k}\sum_{j=1}^{p}(x_{ij} - \bar{x}_{kj})^2\right]$$

bisogna minimizzare *congiuntamente*

- rispetto ai gruppi $G_1, \ldots, G_K$
- ai centrodi $\bar{x}_1, \ldots, \bar{x}_K$

L'algoritmo delle K medie minimizza *localmente* la quantità sopra indicata (non è garantito il minimo globale) minimizzando *in alternanza* rispetto ai gruppi e rispetto ai centrodi

2.1 Algoritmo delle K-medie

1. Si parte con una attribuzione iniziale per $\bar{x}_1, \ldots, \bar{x}_k$
   (e.g. considerando K unità statistiche).
   Si procede iterando ② e ③ fino alla convergenza:

2. Minimizzazione rispetto ai gruppi:
   per $i = 1, \ldots, n$, si individua il centroide più vicino (secondo d^2) all'unità u_i e la si attribuisce al gruppo corrispondente G_k

3. Minimizzazione rispetto ai centroidi:
   per $k = 1, \ldots, K$, si aggiorna il valore del k-simo centroide con la media delle unità del gruppo G_k. Calcolare W.

Si arresta l'algoritmo quando W non cambia rispetto al passo precedente (convergenza)

L'algoritmo definisce una tassellazione di Voronoi in $\mathbb{R}^p$

$$V_k = \{x \in \mathbb{R}^p : d^2(x - \bar{x}_k) \leq d^2(x - \bar{x}_h), h = 1, \ldots, K\}$$

che sono poliedri convessi

2.2 Proprietà dell'algoritmo delle K-medie

- W decresce ad ogni iterazione dell'algoritmo: $W_{i+1} \leq W_i$, dove W_i è W all'iterazione i-sima
- L'algoritmo converge sempre, indipendentemente dall'attribuzione iniziale dei centroidi. Ci mette $\leq K^n$ iterazioni
- I gruppi finali dipendono dall'attribuzione iniziale dei centroidi. Tipicamente si fa girare l'algoritmo più volte inizializzando i centroidi casualmente, e si sceglie il risultato con W minimo
- L'algoritmo non garantisce di minimizzare globalmente W

2.3 Indicazioni sul metodo delle K-medie

- Adatto a scoprire gruppi di forma convessa ma inadatto per gruppi di forma concava
- Il risultato è sensibile alla presenza di valori anomali
- Non è invariate a trasformazioni di scala

2.4 Indice CH

Determinare il numero K di gruppi è in generale un problema piuttosto difficile. Una possibilità è calcolare l'indice CH (Calinski and Harabasz, 1974)

$$\mathrm{CH}(K) = \frac{B(K)/(K-1)}{W(K)/(n-K)}$$

per K che va da 2 a un pre-fissato $K_{\max}$ e si sceglie

$$\hat{K} = \arg \max_{K \in \{2,\ldots,K_{\max}\}} \mathrm{CH}(K)$$

2.5 La *silhouette*

- Determinato, in qualunque modo (non solo con il metodo delle K-medie), un raggruppamento di n unità in K gruppi $G_1, \ldots, G_K$ la *silhouette* è uno strumento per verificare la 'bontà' (coesione interna e separazione esterna) di tale raggruppamento
- Si confronta, per ciascuna osservazione, quanto essa sia vicina al suo gruppo e agli altri.
- La distanza dell'osservazione u_i^* dal gruppo G_k è definita come

$$d(u_i^*, G_k) = \frac{1}{n_k} \sum_{l: u_l \in G_k} d(u_i^*, u_l).$$

- Sia poi G_{k^*} il gruppo in cui è inclusa l'osservazione u_i^* e sia

$$d_0 = \min_{k \neq k^*} d(u_i^*, G_k),$$

  d_0 è la distanza di u_i^* dal gruppo più vicino diverso da quello cui appartiene

- Si confronta d_0 con la distanza dal suo gruppo mediante

$$S(u_{i^*}) = \frac{d_0 - d(u_{i^*}, G_{k^*})}{\max\{d_0, d(u_{i^*}, G_{k^*})\}}.$$

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