Cluster Analysis

Giovanni Corradini

Free University of Bolzano

Contesto Generale

- Il contesto in cui ci muoviamo è quello in cui si dispone di un campione di un certo numero di osservazioni, indicato con n, sui quali vengono misurate/osservate un certo numero di caratteristiche (le variabili), indicato con p, con $p \geq 1$.
- I dati vengono organizzati in una **matrice n x p**, con n righe e p colonne, chiamata anche matrice del disegno.
- I metodi seguenti saranno focalizzati sul contesto in cui le variabili in analisi sono sui numeri reali (o ordinali, come scale Likert), ma verranno anche brevemente esposte le loro estensioni ai casi di variabili qualitative e miste.

Cluster Analysis

- L'analisi dei gruppi (cluster analysis) serve per far emergere dai dati dei **gruppi di osservazioni**, servendosi delle variabili misurate su ogni unità.
- Non si conosce a priori quali siano i gruppi e da quante e quali osservazioni siano formati.
- Ogni gruppo ha la caratteristica di avere osservazioni **simili tra loro al proprio interno, e dissimili da quelle all'interno degli altri gruppi**.

Cluster Analysis

La **similarità** tra le osservazioni si definisce con la distanza fra esse e la distanza più comune è quella euclidea. Indicando con x_i e x_j i vettori relativi all'$i-esima$ e alla $j-esima$ osservazione, la **distanza euclidea** tra x_i e x_j è definita come:

$$d_2(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{h=1}^{p} (x_{ih} - x_{jh})^2}$$

Se le variabili sono su scale diverse bisogna standardizzarle così da renderle confrontabili. Una generica variabile x_k standardizzata è $(x_k - \mu_k)/\sigma_k$, così da renderla a media zero e varianza unitaria.

Cluster Analysis

I due principali metodi di cluster analysis sono quelli partizionali e quelli gerarchici:

- **Partizionali**: si sceglie a priori un numero k di gruppi e in seguito viene determinata una partizione delle osservazioni in k elementi seguendo un determinato algoritmo.
- **Gerarchici**: individuano una sequenza di partizioni nidificate (tante quante sono le osservazioni) e si sceglie un numero k di partizioni, con $1 \leq k \leq n$.

La differenza tra i due metodi è che quelli partizionali prendono in considerazione suddivisioni solo in un numero pari a k (scelto prima dell'attuazione dell'algoritmo), mentre quelli gerarchici possono trovare una partizione in k elementi per qualsiasi k (con $k \in [1, n]$), però la determinazione della partizione dipende anche dalle partizioni precedenti.

Metodi Partizionali

- Il metodo delle **k – medie** è il metodo partizionale più famoso ed utilizzato, ed è basato sulla costruzione di k gruppi (con k scelto a priori), tali che tutte le osservazioni appartenenti ad un certo gruppo siano **più vicine** (**minor distanza** euclidea) al centroide (vettore delle medie) di quel gruppo rispetto ai centroidi di tutti gli altri gruppi.
- La costruzione di tali gruppi è basata su un algoritmo iterativo e pertanto viene esplorata **solo una parte delle partizioni possibili** e quindi non c'è garanzia di ottenere la soluzione ottima in senso assoluto.
- Viene esplorata solo una parte di tutte le possibili partizioni perché crescono con una velocità elevatissima. Per esempio il numero di possibili partizioni di 10 osservazioni in 3 gruppi è 9330, di 20 osservazioni in 3 gruppi è più di mezzo miliardo e di 100 osservazioni in 3 gruppi è nell'ordine di 10^{46} !!

Metodi Partizionali

- Il metodo è computazionalmente efficiente (perchè si esplora solo una piccola parte delle partizioni) però, dato che si basa sulla media (e non su altri indici più robusti), è molto **sensibile alla presenza di valori anomali**.
- Un metodo che risente meno di questo problema è quello dei medoidi (**PAM**): questo è come il k-medie, ma invece che utilizzare come centroide il vettore delle medie (valore artificiale, spesso non presente nei dati), utilizza un valore fra quelli presenti nei dati (il **medoide**), ritenuto centrale rispetto a qualche misura di dissimilarità: tale metodo è quindi più robusto alla presenza di valori anomali ed è anche più interpretabile dato che il medoide è effettivamente presente nei dati. Tuttavia è computazionalmente meno efficiente del k-medie.

Metodi Partizionali

- Se si dispone di dati qualitativi invece che quantitativi, si può utilizzare il metodo delle **k-mode**. Anche questo metodo è come il k-medie, ma invece che utilizzare il vettore delle medie come centroide, utilizza il vettore delle mode e inoltre, invece che utilizzare la distanza euclidea (priva di senso in questo contesto) utilizza delle misure di **dissimilarità** (come quelle di Jaccard o di Hamming).
- Se invece si dispone di **dati misti** (un po' quantitativi e un po' qualitativi), una soluzione può essere quella di utilizzare il PAM sull'indice di dissimilarità di Gower (indice compreso tra 0 e 1) invece che sulle distanze euclidee.

Metodi Gerarchici

I metodi gerarchici individuano una sequenza di partizioni nidificate (tante quante sono le osservazioni), ed in seguito si sceglie la partizione di k elementi desiderata, con $1 \leq k \leq n$.
Tali metodi si dividono in due tipologie:

- **Metodi Agglomerativi**: si parte con tutte le osservazioni in un cluster diverso, per poi agglomerarle in uno unico.
- **Metodi Divisivi**: si parte con tutte le osservazioni in un unico cluster, per poi dividerle ognuna in uno diverso.

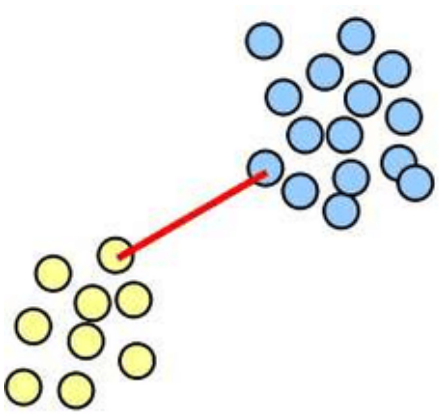
Tali metodi rimangono invariati con variabili qualitative o miste, tuttavia invece che operare il raggruppamento (/divisione) gerarchico sulla matrice delle distanze euclidee, lo si opera sulla matrice delle dissimilarità: nel caso di variabili solamente qualitative si utilizzano le dissimilarità di di Jaccard o di Hamming, mentre se sono miste quelle di Gower.

Metodi Agglomerativi

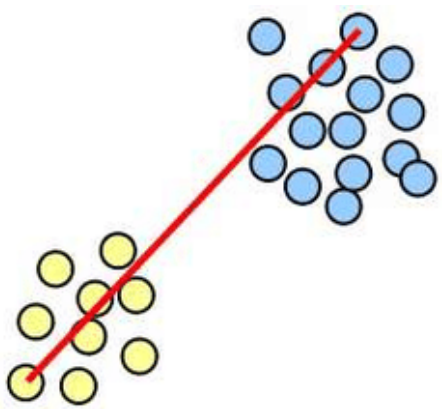
Nei metodi agglomerativi si parte con ogni osservazione in un cluster diverso, e ad ogni passo dell'algoritmo (che in questo caso individua una soluzione unica) si uniscono i due cluster più vicini. Per decidere quali sono i più vicini ci sono vari metodi:

- **Legame singolo**: i due cluster distano tra loro quanto i due elementi più vicini dei due cluster. Con questo tipo di legame si collegano elementi anche distanti tra loro purché tra essi esista una successione di punti che li lega (effetto a catena). Si rischia di legare osservazioni di gruppi diversi.
- **Legame completo**: i due cluster distano tra loro quanto i due elementi più lontani dei due cluster. Con questo tipo di legame si colgono cluster compatti, tendenzialmente di forma circolare. Si rischia di perdere cluster a forma irregolare.

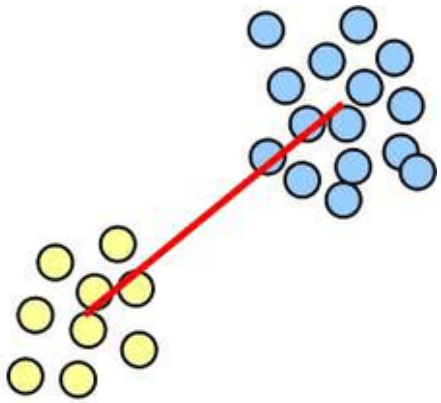
Metodi Agglomerativi



single-link



complete-link



average-link

Metodi Agglomerativi

- **Legame medio**: i due cluster distano tra loro quanto la distanza fra i centroidi dei due cluster. Questo è una via di mezzo tra il legame singolo e quello completo.
- **Criterio di Ward**: unisco i due cluster che creano la configurazione con minor varianza possibile entro i cluster.

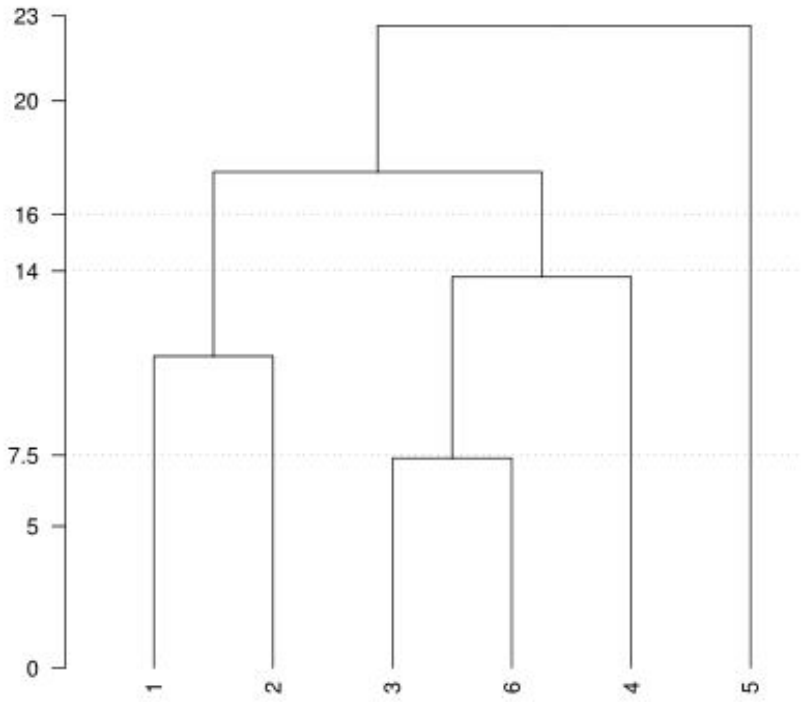
Nessun metodo è superiore ad un altro e scegliendo l'uno o l'altro metodo si ottengono risultati anche molto diversi tra loro. Di caso in caso si sceglie il legame più adeguato in base alla forma dei dati.

Metodi Divisivi

I metodi divisivi funzionano nella maniera opposta rispetto a quelli agglomerativi, ovvero si parte da tutte le osservazioni in un unico cluster per finire con tutte le osservazioni in un cluster diverso. Ad ogni passo dell'algoritmo viene selezionato il cluster con il diametro (dissimilarità) maggiore e questo lo si dividerà in due cluster, tali che ognuno dei due cluster abbia al suo interno la minore dissimilarità media possibile tra le osservazioni.

Metodi Gerarchici

La gerarchia delle partizioni può essere visualizzata tramite un **dendogramma**, che evidenzia i gruppi che si formano per ogni partizione. Una specifica partizione si ottiene tagliando orizzontalmente il dendogramma ad altezza desiderata.



Agglomerativi vs Divisivi

Quando usare i metodi gerarchici agglomerativi e quando quelli divisivi?

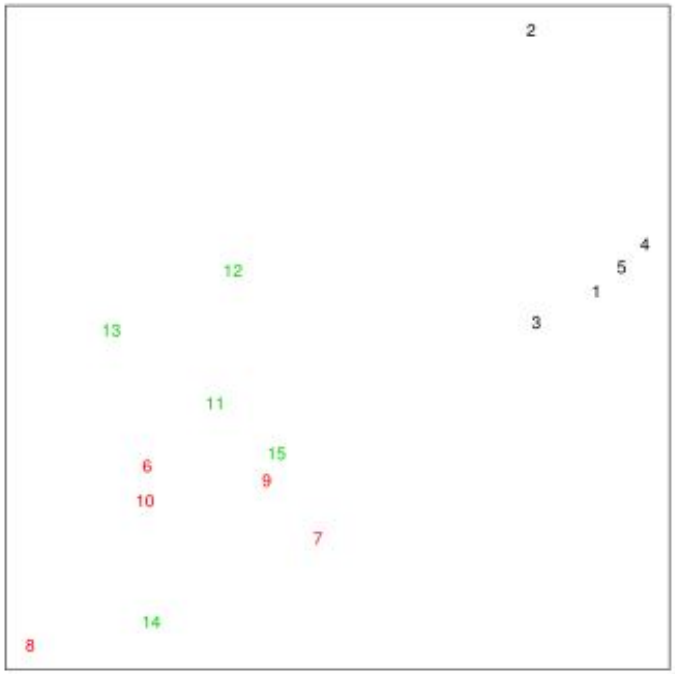
- Dato che i metodi divisivi valutano la struttura globale delle osservazioni (e non i pattern locali come gli agglomerativi) possono essere **più accurati**.
- Però, dato che i divisivi partono con tutte le osservazioni in un unico cluster, hanno un numero maggiore di partizioni possibili e quindi sono **meno efficienti** degli agglomerativi a livello computazionale.
- Inoltre i metodi agglomerativi dispongono, a differenza dei divisivi, di una **varietà di legami**, fatto che li rende adatti in più contesti differenti.

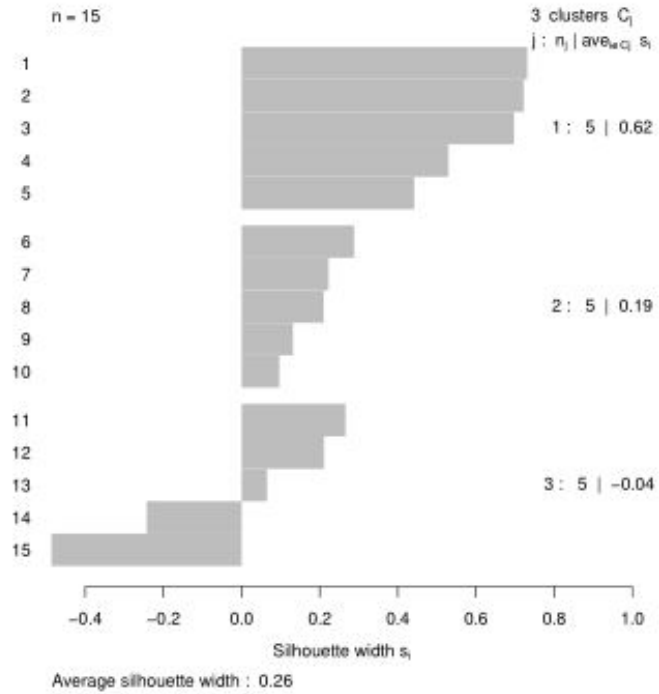
Scelta del numero di gruppi

- Per determinare quanti gruppi scegliere, o a priori nei metodi partizionali o a che punto "fermarsi" (tagliare il dendogramma) nei metodi gerarchici, si utilizza la **silhouette** della partizione, che è un metodo per verificare la "bontà" di tale partizione, ossia in che misura si abbia coesione interna ai gruppi e separazione esterna tra questi.
- Ogni osservazione ha un indice di silhouette che varia tra -1 e 1, dove un valore elevato indica che l'osservazione è "ben abbinata" nel suo gruppo e "mal abbinata" ai gruppi vicini. Se la maggior parte delle osservazioni ha un valore elevato allora la configurazione dei gruppi trovata è adeguata, mentre se il valore è basso la configurazione ha troppi o troppi pochi gruppi.

Scelta del numero di gruppi

Per trovare il numero di gruppi adeguato (o a priori nei metodi partizionali, o dove "tagliare" nei metodi gerarchici) si analizza il grafico (il valore medio) della silhouette al variare della numerosità dei gruppi, e si sceglie quindi la configurazione con il **valore di silhouette più elevato**.





Partizionali vs Gerarchici

Quando usare i metodi gerarchici e quando quelli partizionali?

- Se **conosco a priori il numero di cluster** che voglio individuare (ma non so quali siano) allora i metodi partizionali sono più appropriati. Inoltre se ho un dataset con molte osservazioni i metodi partizionali sono più **efficienti** a livello computazionale (dato che viene trovato un massimo locale e non globale).
- Se **non conosco a priori il numero di cluster** allora con i gerarchici riesco a capire più facilmente il numero di cluster adeguato. I metodi gerarchici sono più **informativi** in quanto si vede come sono stati creati i cluster. Inoltre ci sono contesti (come la tassonomia) in cui la struttura di fondo dei dati è organizzata gerarchicamente, e in tal caso sicuramente i metodi gerarchici sono i più adeguati.