Capitolo 8

Materiale didattico integrativo

Approfondimenti sull'analisi cluster

Sommario

1. Altri algoritmi gerarchici agglomerativi 1
2. Metodi non gerarchici. Il metodo *k-means* 2
3. Sul funzionamento dei vari metodi di cluster 3

1. Altri algoritmi gerarchici agglomerativi

Esistono altri algoritmi agglomerativi che, per motivi di spazio, non sono stati considerati nel testo.

I più noti sono:

- il metodo del **centroide**
- il metodo della **mediana**
- il metodo della **distanza media ponderata** (o metodo della **media ponderata**)

Il metodo del **centroide** calcola le distanze fra gruppi applicando ai centroidi di gruppo (le medie delle p variabili) la formula della distanza scelta. Vengono uniti i due gruppi che hanno la più piccola distanza fra le medie. Il centroide del nuovo gruppo è la media delle variabili delle unità appartenenti al gruppo stesso.

Se si uniscono i gruppi G_1 e G_2 di dimensione n_1 e n_2 e con centroidi C_1 e C_2, il centroide C_3 del nuovo gruppo è calcolato come:

$$C_3 = \frac{C_1 n_1 + C_2 n_2}{n_1 + n_2}$$

Il valore di C_3 sarà influenzato maggiormente dal valore del centroide del gruppo più numeroso.

Questo metodo, non è basato sul confronto fra le distanze fra singole unità particolari (come ad esempio il criterio del legame singolo, del legame completo) ma tra distanze fra indici caratteristici globali dei gruppi via via formatisi.

Il metodo della **mediana** è simile a quello del centroide ma da questo si differenzia per come viene calcolato l'indice caratteristico globale del gruppo (che non è più un centroide), una volta che il gruppo si è formato. Invece di calcolare la media delle unità appartenenti al gruppo, viene calcolata la semisomma degli indici caratteristici dei due gruppi che si sono uniti. L'indice rappresentativo C_3^* del nuovo gruppo formato dall'unione dei gruppi G_1 e G_2 sarà allora:

$$C_3^* = \frac{C_1^* + C_2^*}{2}$$

Questo metodo elimina l'effetto della diversa dimensione dei gruppi e tende così a favorire la formazione di gruppi di dimensioni simili perché l'indice caratteristico del nuovo gruppo non sarà maggiormente influenzato dal valore dell'indice del gruppo più numeroso.

Secondo il metodo della **distanza media ponderata** (o della media ponderata), la distanza fra due gruppi è data dalla somma delle distanza fra le unità di un gruppo e quelle dell'altro gruppo (invece della media delle distanze che è usata nel criterio della distanza media; v. Paragrafo 8.5 del testo). Questo metodo, eliminando l'effetto della diversa dimensione dei gruppi, tende a unire gruppi di dimensioni più piccole e quindi tende a favorire la formazione di gruppi di dimensioni simili.

2. Metodi non gerarchici. Il metodo *k-means*

La procedura di base con cui procede la cluster gerarchica agglomerativa inizia col calcolo delle dissimilarità o distanze fra unità e termina con la costruzione del dendrogramma che evidenzia i risultati di ogni passo in cui viene ottimizzata una certa funzione obiettivo (si minimizza la dstanza o l'indice di distanza): gli oggetti con la minima distanza vengono uniti fra loro. Le differenze fra i vari criteri risiedono nel diverso modo di definizione della distanza fra unità, unità e gruppo e fra gruppi.

Il dendrogramma, che descrive la procedura aggregativa, non stabilisce automaticamente il numero di gruppi ma una serie progressiva di possibili raggruppamenti. Quella definitiva – che sarà una partizione delle unità – viene scelta tagliando l'albero (operazione di *pruning*) aiutandosi anche con lo *screeplot*.

Anche coi metodi non gerarchici l'attribuzione di una unità ad un gruppo avviene ottimizzando una funzione obiettivo (minimizzando la distanza o un indice di distanza). Tali criteri tuttavia si differenziano da quelli gerarchici sia perché forniscono direttamente, nella generalità dei casi, una sola partizione sia perché consentono lo spostamento delle unità ai gruppi, rendendo possibile la correzione di soluzioni insoddisfacenti. In genere però necessitano di conoscere preliminarmente il numero di gruppi.

Nell'ambito dei metodi gerarchici, il termine *k-means* è stato introdotto da MacQueen nel 1967 che propose una variante del metodo proposto da Forgy nel 1965. Col tempo, il termine *k-means* è andato a rappresentare vari algoritmi non gerarchici di identificazione di gruppi disgiunti e esaustivi, algoritmi che si differenziano fra loro per alcuni aspetti tecnici pur rispettando la stessa logica di fondo.

L'algoritmo di Forgy, proposto nel 1965 segue una procedura molto semplice che riassumiamo nelle seguenti fasi.

1) Si stabilisce il numero *k* dei gruppi.

2) Si individua no *k* centroidi di partenza.

3) Si attribuisce ogni unità di osservazione al gruppo più vicino calcolando la distanza fra l'unità e il centroide del gruppo.

4) Quando tutte le unità sono state assegnata ad un gruppo, si calcolano i nuovi centroidi.

5) Ripetizione delle fasi 3 e 4 fino a quando nessuna unità cambia il gruppo di assegnazione.

L'algoritmo di MacQueen proposto nel 1967 e denominato originariamente proprio criterio *k-means*, processa le unità in modo sequenziale.

1) Si stabilisce il numero dei gruppi *k*.

2) Si individuano *k* centroidi iniziali (le prime *k* unità).

3) In modo sequenziale ogni unità viene assegnata al gruppo col centroide più vicino e ad ogni singola assegnazione si aggiornano i centroidi.

4) Si ripete la fase 3 fino a quando nessuna unità cambia il gruppo di assegnazione.

In genere, i metodi non gerarchici di questo tipo si possono differenziare per almeno tre aspetti.

a) Le modalità di scelta dei centroidi di partenza. Si può ad esempio individuare una partizione iniziale aggregando in qualche modo le unità in *k* gruppi e poi calcolare i centroidi dei gruppi così formati, o stabilendo, con qualche criterio, *k* valori che rappresentano i centroidi iniziali.

b) Il momento di aggiornamento dei valori dei centroidi: alla fine dell'assegnazione di tutte le unità (Forgy) o via via che una unità è assegnata ad un gruppo (MacQueen).

c) La funzione di distanza adottata. In genere è la distanza euclidea o il suo quadrato. Se si usa il quadrato della distanza euclidea, entrambi i criteri visti, basati sull'allocazione delle unità al gruppo col centroide più vicino, tendono a minimizzare (ad ogni passo della procedura) la devianza interna ai gruppi (si veda al proposito la formula (17) del Paragrafo 8.5.3 del testo). Ad ogni modo, qualunque sia la funzione di distanza, poiché l'assegnazione delle unità avviene in base alla distanza dal centroide di gruppo, i due metodi tendono a individuare gruppi sferici.

3. Sul funzionamento dei vari metodi di cluster

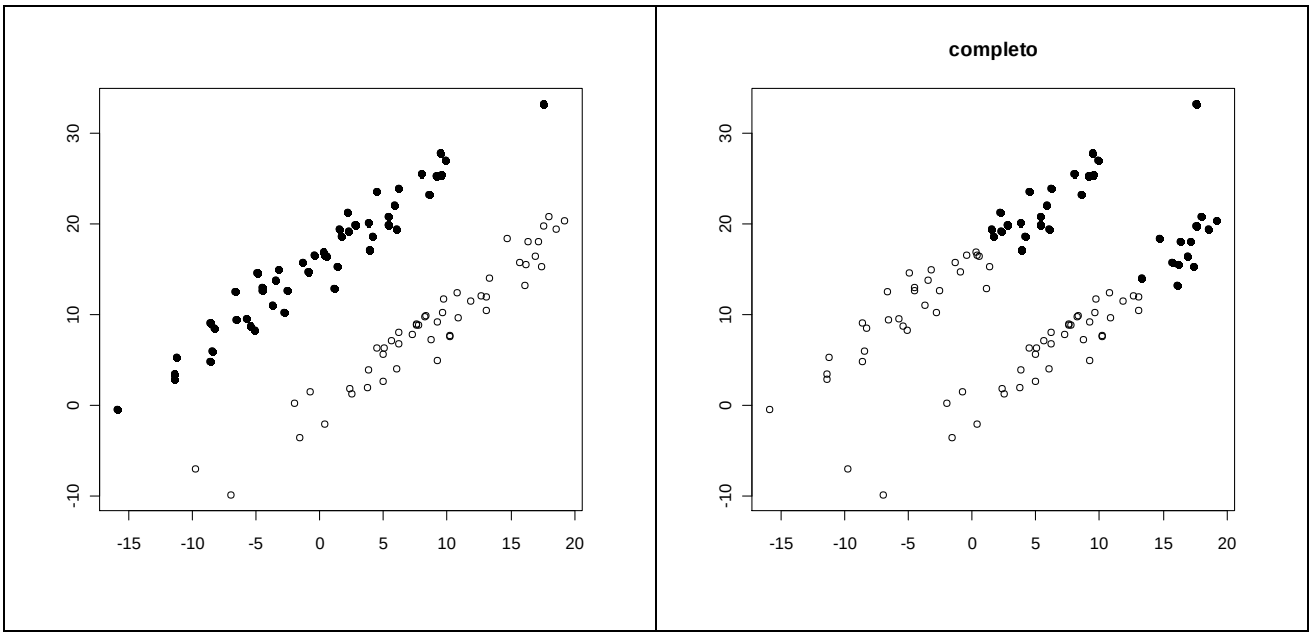
I criteri gerarchici agglomerativi presentati qui e nel testo sono quelli maggiormente impiegati nella pratica e previsti da tutti i software statistici. Per quanto riguarda le possibilità di utilizzo dei vari criteri gerarchici agglomerativi, non esiste un criterio che va bene in tutte le occasioni.

Dal punto di vista puramente teorico, la preferenza dovrebbe andare ai criteri del legame singolo o del legame completo che non introducono artifici né attraverso sintesi di distanze (come il criterio della distanza media e quello della distanza ponderata) né col calcolo di indici caratteristici di gruppo ma, ad ogni passo della procedura, la matrice delle distanza è sempre composta da distanze fra unità (si veda anche l'esempio riportato nel testo).

Vari studi di simulazione hanno confrontato il funzionamento dei metodi di cluster, gerarchici e non gerarchici. In molti contributi, i migliori algoritmi gerarchici sono risultati quello della distanza media e quello di Ward. Il peggiore: il metodo del legame singolo.

Il motivo principale di questi risultati risiede nel fatto che la maggior parte dei lavori ha simulato gruppi sferici che vengono facilmente individuati da quasi tutti i metodi, soprattutto da quello di Ward e anche dal *k-means*. Tuttavia questi metodi non sono idonei a trovare "gruppi allungati" .

L'esempio con 2 variabili sotto riportato mostra due gruppi allungati (figura a sinistra) che non vengono identificati direttamente dai metodi introdotti. A titolo di esempio, nella figura a destra, mostriamo il risultato del metodo del legame completo applicato ai dati originari.



La situazione di gruppi allungati è in genere determinata da variabili correlate e/o aventi diversa variabilità. La soluzione in questo caso è quella di trasformare le variabili originari mediante l'analisi delle componenti principali (meglio se su variabili standardizzate) e poi applicare l'analisi cluster alle componenti principali standardizzate. Il ricorso a forme di standardizzazione o all'ACP tuttavia non è *indolore* come è stato accennato nel testo (v. Esempio 8.4 del Capitolo 8).

Nell'esempio della figura, si riesce a identificare perfettamente i due gruppi seguendo le fasi sotto:

1) si ricavano le componenti principali dai dati standardizzati (matrice di correlazione);
2) si standardizzano le componenti principali;
3) si applica il metodo di Ward alle componenti principali standardizzate.

Come metodo esplorativo in sé, l'analisi cluster deve essere usata soprattutto per semplificare e descrivere i dati di osservazione. Devono essere prevalentemente ragioni di ordine pratico a indirizzare verso il criterio più appropriato, da valutare in base all'interpretazione dei gruppi individuati e anche alla luce delle conoscenze che si hanno sul fenomeno indagato.

Le ragioni di ordine pratico portano spesso ad eliminare il metodo del legame singolo.