

una matrice di prossimità o di distanza che contiene cioè le informazioni relative alla distanza o alla similarità di ciascun oggetto rispetto agli altri. Successivamente si usa l'informazione contenuta in questa matrice per unire gli oggetti nei gruppi. Nella matrice di prossimità sia le righe che le colonne sono rappresentate dagli oggetti, e i valori contenuti nella matrice indicano la similarità o la differenza fra ogni coppia di osservazioni.

A prima vista potrebbe sfuggire la differenza fra una matrice di correlazione e una matrice di prossimità, ma la differenza è sostanziale. Infatti, mentre la prima contiene similarità fra variabili ($x_1; x_2$), una matrice di prossimità contiene similarità fra osservazioni ($o_1; o_2$).

I metodi di Cluster Analysis possono essere distinti in due grosse famiglie: agglomerative e divisive.

Metodi agglomerativi. I metodi di cluster analysis di questo tipo iniziano con tanti gruppi quanti sono gli oggetti e ad ogni passo due clusters sono fusi; questo tipo di operazione va avanti fino a quando non si crea un unico grosso cluster. È il ricercatore che deve decidere a quale passo fermare il raggruppamento e in questo modo decidere quanti gruppi ottenere.

Metodi divisivi. I metodi di cluster analysis di questo tipo lavorano in maniera opposta rispetto ai metodi agglomerativi. Inizialmente tutte le osservazioni fanno parte di un unico grosso gruppo. Ad ogni operazione, il gruppo iniziale viene scisso fino a che tutte le osservazioni non costituiscono un cluster a se. Anche qui, è il ricercatore che deve decidere il numero di cluster ottimale.

A questo punto possiamo individuare alcuni passaggi nella Cluster Analysis di tipo agglomerativo:

1. Bisogna decidere quali dati costituiscono i nostri casi.
2. Viene calcolata la distanza o la similarità per ogni coppia degli N oggetti da classificare (matrice di distanza o similarità $N \times N$). Per ottenere la matrice di distanza sarà necessario computare $N \times (N-1) / 2$ valori distinti.
3. Si considerano inizialmente N gruppi, ciascuno costituito da un singolo oggetto.
4. Si cerca nella matrice delle distanze per trovare

i due cluster che risultano tra loro più vicini.

5. Si fondono in un unico cluster i gruppi che risultano più vicini.
6. Si calcola la distanza fra questo nuovo cluster e tutti gli altri (che contengono ciascuno un solo elemento).
7. Si ripete il processo di fusione $N-1$ volte a partire dal passo numero 4, finché tutti i gruppi non sono fusi in un unico gruppo.

Le diverse tecniche gerarchico-agglomerative si differenziano per le modalità secondo cui effettuano il passo 6, cioè per i diversi criteri con i quali avviene la fusione dei gruppi. Nel punto 2 si diceva che per ottenere la matrice di prossimità sarà necessario computare $N \times (N-1) / 2$ valori distinti. Per questo motivo se si cerca di classificare con le tecniche gerarchico-agglomerative insieme composti da più di qualche centinaio oggetti, si incontrano molte difficoltà.

Scelta dell'algoritmo di classificazione

Definito il criterio per la classificazione, ossia stabilito il tipo di misura di somiglianza/ diversità da adottare, si potrebbe pensare che la procedura ottimale per procedere, sia quella di esaminare tutte le possibili partizioni dell'insieme degli oggetti, per limitarsi a valutare poi solamente quella che nel modo migliore soddisfa il criterio prescelto. Tuttavia, quando il numero degli oggetti da classificare è maggiore delle 15-20 unità, questa operazione diventa praticamente impossibile, in quanto il numero delle partizioni possibili diventa enorme. Pertanto le diverse procedure di analisi dei gruppi si presentano come strategie che esplorano solo un limitato sottoinsieme di tutte le possibili partizioni (Massart *et al.*, 1988). Bisogna a questo punto scegliere il "clustering algorithm" ossia l'insieme delle regole che stabiliscono come le distanze tra i gruppi vanno misurate. I criteri usati da questi algoritmi sono differenti, e pertanto dagli stessi dati possono essere ottenute differenti classificazioni, anche qualora venisse usata la stessa misura di similarità/ distanza. È importante tenere sempre presente che sebbene la Cluster Analysis rientri nei metodi oggettivi di analisi, esistono degli elementi di soggettività nella scelta del metodo.

In SPSS sono presenti cinque differenti