

Omogeneizzazione delle scale di misura

Ogni variabile può avere estensioni di scala differente. Affinché non vengano introdotte distorsioni nel calcolo delle misure di similarità dovute alle disomogeneità delle scale, è necessario operare una trasformazione delle scale originarie.

Un modo per effettuare questa trasformazione è quello di dividere i valori di ciascuna variabile per il suo valore medio. In questo modo, ciascuna variabile sarà trasformata in un numero puro, con media uguale a 1 (Massart *et al.*, 1988).

Il modo più semplice per risolvere questo inconveniente è procedere con la standardizzazione delle variabili, che permette appunto di dare loro uguale unità di misura. Si opera sottraendo a ciascun valore il valore medio della variabile stessa e si divide il risultato ottenuto per la deviazione standard. Tutte le variabili così ottenute avranno come media il valore 0 e come varianza 1.

$$z_i = \frac{x_i - \bar{x}}{s}$$

dove s è la deviazione standard della variabile x .

Ossia:

$$\text{variabile standardizzata} = \left(\frac{\text{variabile osservata} - \text{media variabile osservata}}{\text{deviazione standard della variabile osservata}} \right)$$

In questo modo si perde naturalmente l'informazione relativa al diverso grado di variabilità delle variabili stesse, informazione che in alcuni casi può essere rilevante per il ricercatore.

Misure di somiglianza e di diversità

Uno degli approcci più usati per accostarsi al problema della somiglianza/ diversità degli oggetti si basa sul concetto di distanza, naturalmente pertinente solo nel caso di variabili cardinali. La classificazione che si ottiene applicando la Cluster Analysis è influenzata dal particolare metodo usato.

Esistono diversi tipi di distanza che è possibile calcolare:

$$\text{distanza euclidea: } D_{ij} = \left[\sum_{k=1}^v (X_{ik} - X_{jk})^2 \right]^{1/2}$$

Ove la sommatoria è estesa a tutte le variabili utilizzate.

distanza assoluta:

$$D_{ij} = \sum_{k=1}^v |X_{ik} - X_{jk}|$$

La distanza Euclidea e la distanza Assoluta possono essere considerate due casi particolari della distanza di Minkovski, definibile in generale con la formula:

distanza di Minkovski:

$$D_{ij} = \left[\sum_{k=1}^v (X_{ik} - X_{jk})^r \right]^{1/r}$$

La distanza euclidea corrisponde al caso in cui r è uguale a 2, quella assoluta al caso in cui r è uguale a 1.

Distanza di Mahalanobis:

$$D_{ij} = \left[(X_i - X_j) Z^{-1} (X_i - X_j) \right]^{1/2}$$

dove X_i ed X_j sono due vettori riga che contengono i valori relativi all' i -esimo caso e al j -esimo caso, Z è la matrice delle varianze-covarianze e Z^{-1} è la matrice inversa della matrice varianze-covarianze. La distanza di Mahalanobis coincide con la distanza Euclidea nel caso di assenza di correlazione fra qualsiasi coppia di variabili.

Scelta della tecnica di classificazione

Le più diffuse tecniche di classificazione sono suddivise in due grandi categorie:

Le tecniche di classificazione gerarchica che producono una sequenza di $N-1$ partizioni disposte in ordine gerarchico.

Le tecniche di classificazione non gerarchica o a partizioni ripetute che permettono di determinare direttamente una partizione dei casi in rapporto a uno specifico criterio prefissato dal ricercatore.

In funzione dei dati, si sceglierà se applicare l'uno o l'altra metodologia. In questa ricerca si è applicata la prima che si cercherà di approfondirla.

Le tecniche di classificazione gerarchica (hierarchical cluster)

In qualsiasi algoritmo che ambisce alla costruzione di una classificazione, il punto di partenza è costituito da una matrice dati Casi $\times$ Variabili ($C \times V$) dove cioè sulle righe si hanno i casi, mentre sulle colonne le variabili. Il primo passo di una tecnica di analisi dei gruppi è quello di trasformare una matrice $C \times V$, in una matrice $N \times N$, ossia in