Segmentazione delle famiglie in gruppi sociali

La definizione dei gruppi sociali in funzione del reddito è stata condotta mediante l'ausilio di tecniche statistiche non parametriche di segmentazione gerarchica o CART (Classification and Regression Trees) che possono essere convenientemente utilizzate quando le ipotesi teoriche e distributive dei metodi di classificazione classici non sono sostenibili o sono difficilmente verificabili.[1]

Tali numerose tecniche si propongono di individuare classi latenti in basi di dati particolarmente numerosi utilizzando una variabile dipendente Y e un insieme di variabili esplicative X1, …, Xm, al contrario dei metodi di *clustering* che costruiscono gruppi di unità partendo da un insieme indistinto di variabili.

In particolare, la segmentazione gerarchica è una procedura iterativa attraverso la quale l'insieme delle n unità statistiche viene suddiviso progressivamente in una serie di sottogruppi disgiunti, basati sulle modalità di una o più variabili esplicative, maggiormente omogenei rispetto alla variabile dipendente. Pertanto, la segmentazione fornisce una successione gerarchica di partizioni dell'insieme delle n unità (nodi) ottenuta con un criterio scissorio o top down. Il risultato è rappresentato mediante una struttura grafica detta 'albero decisionale' o 'albero di classificazione'.

La procedura adottata è una variante del metodo CHAID (Chi-square Automatic Interaction Detection), per variabili dipendenti quantitative, con suddivisione dei nodi di tipo binario.

Il metodo determina, a ogni passo, la migliore suddivisione delle unità statistiche per ogni predittore, fondendo iterativamente coppie simili di unità, finché non rimane una sola coppia di gruppi. Tutte le coppie vengono poi confrontate tra loro, al fine di scegliere la migliore in termini di omogeneità dei gruppi ottenuti. La scelta si basa sul test statistico F di ANOVA (Analysis of Variance) che ha lo scopo di valutare se le medie di Y, nei gruppi, sono uguali (ipotesi nulla). I gruppi ottenuti sono stati 'sfrondati' (pruning), in modo da minimizzare la complessità a parità di potere discriminatorio, fino ad ottenere nove gruppi sociali.

La connessione tra le diverse variabili esplicative e la classificazione delle famiglie è stata misurata mediante l'indice V di Cramer. Questo indice varia tra 0 e 1, assume il valore 0 in caso di indipendenza e il valore 1 in caso di massima dipendenza. La lista delle variabili esplicative per valori decrescenti del V di Cramer, tra parentesi, è la seguente:

- presenza di stranieri (0,858);
- situazione professionale della persona di riferimento (0,609);
- titolo di studio della persona di riferimento (0,548);
- sesso della persona di riferimento (0,321);
- età della persona di riferimento (0,304);
- numero di componenti (0,284);
- tipo di comune di residenza (0,084).

Per saperne di più

S. Zani, A. Cerioli (2007). Analisi dei dati e data mining per le decisioni aziendali. Giuffré Editore, Milano.

1 Zani e Cerioli (2007).