On the Analysis of Symbolic Data

Paula Brito

Faculdade de Economia/NIAAD-LIACC, Universidade do Porto
Rua Dr. Roberto Frias, 4200-464 Porto, Portugal

Abstract. Symbolic data extend the classical tabular model, where each individual, represented by a row, takes exactly one value for each variable represented by a column, by allowing multiple, possibly weighted, values for each variable. New variable types - interval, categorical multi-valued and modal variables - have been introduced, which allow representing variability and/or uncertainty inherent to the data.

But are we still in the same framework when we allow for the variables to take multiple values? Are the definitions of basic notions still so straightforward? What properties remain valid?

In this talk we will discuss some issues that arise when trying to apply classical data analysis techniques to symbolic data. The central question of the evaluation of dispersion, and the consequences of different possible choices in the design of multivariate methods, will be addressed.

Dispersion is a key issue in clustering, since the result of any clustering method depends heavily on the scales used for the variables; natural clustering structures can sometimes only be detected after an appropriate rescaling of variables. The standardization problem has been addressed by De Carvalho, Brito & Bock, and three standardization techniques for interval-type variables have been proposed.

Furthermore, many exploratory multivariate methodologies rely heavily on the notion of linear combination and on the properties of dispersion measures under linear transformations. This problem has been addressed in a recent work of Duarte Silva & Brito in the context of linear discriminant analysis of interval data.

Keywords

SYMBOLIC DATA ANALYSIS, INTERVAL DATA, STANDARDIZATION, CLUSTERING, DISCRIMINANT ANALYSIS