DIPARTIMENTO DI SCIENZE STATISTICHE

Seminar

MULTILEVEL TYPOLOGIES: CLASSIC AND SYMBOLIC DATA APPROACHES

JOSE G. DIAS

ISCTE – INSTITUTO UNIVERSITARIO DE LISBOA

December 4, 2017

10.30

Aula Cucconi

Abstract
www.stat.unipd.it/fare-ricerca/seminari

MULTILEVEL TYPOLOGIES: CLASSIC AND SYMBOLIC DATA APPROACHES

JOSÉ G. DIAS

ISCTE-INSTITUTO UNIVERSITÁRIO DE LISBOA

Abstract

The increasing amount of available data on the Internet together with new technologies that allow linking and integrating data from heterogeneous sources have put enormous pressure on the development of new analytic frameworks. New types of data may have many distinct characteristics that result from the integration of different layers of hierarchically structured complex systems from micro to macro levels, i.e., observed units are nested within units of higher levels. Classical examples are patients nested within hospitals, students within classrooms, children within families, or employees within firms. Other examples may deal with space and time dependencies. These multilevel structures have been extensively researched in the context of statistical modelling.

A typical problem facing researchers using these complex data structures is to find a typology of individual observations or units. If these units were randomly selected from a given population, then traditional clustering (probabilistic or heuristic-driven algorithms) can be applied. In case these units are organized within upper-level units, the independence assumption is violated and results do not take the heterogeneity at the upper-level into account. Alternatively, symbolic data tend to be more focused on the upper-level and aggregate data at lower level. For instance, clustering schools may result from the analysis of the distribution of characteristics of students within. This aggregation ignores the nested structure and heterogeneity at the individual level.

This paper discusses the clustering of nested structures. We compare multilevel latent class models that take the nested structure into account with alterative solutions. In particular, we emphasize the advantage of clustering upper-level units without aggregating lower-level units. Results using synthetic and empirical data provide important recommendations on how to handle these complex data sets.