Challenges in Classification and Regression in the Era of Big Data

Maurizio Vichi
Sapienza University of Rome, Dpt. Statistical Sciences, Rome, Italy – Maurizio.vichi@uniroma1.it

Big Data frequently describe complex economic, social and demographic phenomena that manifest themselves on individuals (units, objects, sites, with a spatial location), by means of a set of variables differing over individuals and showing both a diffusion over space and an evolution over time. *Big Data* of these type are generally rearranged in large and complex data structures such as three or high dimensions arrays (data (iper)-cubes), corresponding, at least, to a huge number of statistical units (rows) with a spatial location, variables, (columns), times (tubes). These data are generally statistically analysed to describe different relations between, objects (spatial correlation), between variables (cross-sectional correlation) and between times (time series correlation) with the objective to synthesize the relevant information and describe the appropriate relations.
The high dimensionality of the data induces the use of the Latin paradigm "*Divide et impera*" in order to recursively breaking down the statistical analysis into two or more "homogeneous" and less high-dimensional sub-analyses, until these become simple enough to be analysed directly. The solutions to the sub-analyses are then used to give a solution to the original big data analysis.
Such divide and conquer paradigm is strongly linked to the idea of including a clustering model in the statistical analysis of the big data. A modelling approach is proposed for different statistical analysis such as regression and structural equation modelling.

Keywords: Divide et impera paradigm; clustering; regression; Structural Equation Modelling