Symposium on Big Data in Finance, Retail and Commerce: Statistical and Computational Challenges
Lisbon, 2-3 November 2017, Portugal

Analysing (Micro) Big Data at the Macro Level: The Case of Risk Categories of Loan Data

Paula Brito (Speaker)
Fac. Economia & LIAAD-INESC TEC, Univ. Porto, Portugal, mpbrito@fep.up.pt

A. Pedro Duarte Silva
Católica Porto Business School, & CEGE, Univ. Católica Portuguesa, Porto, Portugal, psilva@porto.ucp.pt

Abstract

In Data Mining, Multivariate Data Analysis and classical Statistics, data is usually represented in a data array where each row represents a "case", or "individual", for which one single value is recorded for each numerical or categorical variable (in columns). This representation model is however restricted when the data to be analysed comprises variability, that is the case when the entities under analysis are not single elements, but groups formed on the basis of some given common properties. Then, for each descriptive variable, the observed variability inherent to each group should be taken into account, to avoid an important loss of pertinent information. This is relevant in Data Mining applications where huge sets of data are collected, but data should be analysed at a higher level - e.g. take the case of large department stores, which record data on each purchase made (amount spent, items purchased, etc.), but where the focus does not lie on individual purchases but rather on consumer behaviour, and therefore information about the purchases of each client, or specific groups of clients, must be somehow aggregated. Symbolic Data Analysis [1] provides a framework for the representation and analysis of such data, which comprehends inherent variability. New variable types have been introduced whose realizations are not single real values or categories, but sets, intervals, or, more generally, distributions over a given domain. Methods for the (multivariate) analysis of such symbolic data have been developed which allow taking into account the variability expressed in the data representation.

We focus here on interval-valued data, i.e., where for each entity under analysis

an interval is recorded. In [2] a parametric modelling for interval data, assuming multivariate Normal or Skew-Normal distributions for the MidPoints and Log-Ranges of the interval variables is proposed. The intrinsic nature of the interval variables leads to special structures of the variance-covariance matrix. This modelling allows for multivariate parametric analysis ; in particular M(ANOVA) [2], discriminant analysis [4], model-based clustering [3] and outlier detection [5] have been addressed under this framework.

In this talk we analyse a large dataset of loan data, with more than half a million records, where our main units of interest are risk categories. The microdata have thus been aggregated in the form of intervals, for the variables considered. Multivariate analysis of the resulting interval-valued data reveal insights that we not easily captured at the original microdata level.

Keywords: Interval Data ; Multivariate Data Analysis ; Symbolic Data Analysis

References

[1] Brito, P. (2014). Symbolic data analysis: another look at the interaction of Data Mining and Statistics. *WIREs Data Mining and Knowledge Discovery*, 4 (4), 281–295.

[2] Brito, P. and Duarte Silva, A.P. (2012). Modelling interval data with Normal and Skew-Normal distributions. *Journal of Applied Statistics*, 39 (1), 3–20.

[3] Brito, P., Duarte Silva, A. P. and Dias, J. G. (2015). Probabilistic clustering of interval data. *Intelligent Data Analysis*, 19 (2), 293–313.

[4] Duarte Silva, A.P. and Brito, P. (2015). Discriminant analysis of interval data: an assessment of parametric and distance-based approaches. *Journal of Classification*, 32 (3), 516–541.

[5] Duarte Silva, A. P., Filzmoser, P. and Brito, P. Outlier detection in interval data. *Under revision.*