ERATO セミナ 2010 - No. 35

Introduction to Symbolic Data Analysis - An interaction movement between statistics and data processing -

Dr. Paula Brito
Statistics and Data Analysis at the Faculty of Economics
The University of Porto

2011/2/7

概要

In classical data analysis, data is represented in a $n \times p$ matrix where n individuals (in rows) take exactly one value for each variable (in columns). However, this model is too restrictive to represent complex data, which may comprehend variability and/or uncertainty. The need to consider data that contain information which cannot be represented within the classical framewok lead to the development of Symbolic Data Analysis. Symbolic data extend the classical model, by allowing multiple, possibly weighted, values for each variable. New variable types have therefore been introduced, which allow us to represent variability and/or uncertainty inherent to the data: multi-valued variables, interval variables and modal variables. This approach is of particular and growing interest in the analysis of huge sets of data, recorded in very large databases, when the units of interest are not the individual records (the microdata), but rather some second-level entities. For instance, in a database of credit card purchases, we are probably more interested in describing the behaviour of some person (or even some pre-defined class or group of persons) rather than each purchase by itself. By aggregating the purchase data for each person (or group), we obtain the information of interest; the observed variability for each person or within each group is preserved, and ith is of utmost importance. But are we still in the same framework when we allow for the variables to take multiple values? Are the definitions of basic statistical 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 work of Duarte Silva & Brito in the context of linear discriminant analysis of interval data. Different approaches have been considered by various authors to address these and other questions and to propose a symbolic counterpart of statistical multivariate data analysis methods. As we can see, in recent years there has been a growing literature proposing methodologies for the analysis of symbolic data. However, most existing methods take a non-parametric descriptive approach. In a recent work, Brito and Duarte Silva focus on the analysis of interval data, for which probabilistic models are proposed and used. For modelling purposes, a parameterization consisting in representing each interval $I_{ij} = [l_{ij}, u_{ij}]$ by its midpoint $c_{ij} = (l_{ij} + u_{ij})/2$ and range $r_{ij} = u_{ij} - l_{ij}$ is adopted. The approach consists in modelling each pdimensional interval vector by a 2p-dimensional Normal or Skew-Normal distribution for the interval midpoints and log-ranges. One advantage of the Normal model is its analytical tractability and the possibility of straightforward applications of classical inference methods. On the other hand, the Skew-Normal model offers greater flexibility for the shape of the distributions. In the most general formulation we allow for non-zero correlations among all midpoints and log-ranges, but other particular cases of interest may also be taken into account. The proposed modelling may be applied to multivariate methodologies where a parametric distribution is to be assumed. Firstly, this modelling is employed in the context of (M)ANOVA techniques. This allows, in particular, assessing the relevance of different (interval) variables for a given partition on interval data. Secondly, a model-based clustering method (without fixing in advance the number of clusters), is developed, where a mixture distribution approach is followed, parameters and clusters are determined by maximum likelihood via the EM algorithm. This framework may be extended to other statistical methodologies, opening the way to inference approaches for symbolic data. In a most resent approach, quantile representation (Ichino, 2008) provides a common framework to represent symbolic data described by variables of different types. The principle is to express the observed variable values by some predefined quantiles of the underlying distribution. In the interval variable case, a distribution is assumed within each observed interval, e.g. uniform (Bertrand and Goupil, 2000) ; for a histogram-valued variable, quantiles of any histogram may be obtained by simply interpolation, assuming a uniform distribution in each class (bid); for categorical multi-valued variables, quantiles are determined from the ranking defined on

the categories based on their frequencies. When quartiles are chosen, the representation for each variable is defined by the 5-tuple (Min, Q1, Q2, Q3, Max). This common representation then allows for a unified analysis of the data set, taking all variables simultaneously into account. In a numerical clustering context, the Ichino-Yaguchi dissimilarity is used to compare data units; hierarchical and pyramidal models, with several aggregation indices, may be applied and clusters are formed on the basis of quantile proximity. We also focus on a conceptual clustering approach. In this case, clusters are represented, for each variable, by a mixture of the quantile-distributions of the merged clusters and then compared on the basis of the current quantile representation. The proposed hierarchical/pyramidal clustering model follows a bottom-up approach; at each step, the algorithm selects the two clusters with closest quantile representation to be merged. The newly formed cluster is then represented according to the same model, i.e., a quantile representation for the new cluster is determined from the uniform mixture cumulative distribution. Much remains however to be done. We have just rather briefly discussed some of the issues that arise when we leave the classical data framework and allow for more complex variable types. Usual properties, generally taken for granted, often do not apply any longer, and new concepts much be put forward. Among these, parametric statistical analysis is an important challenge. A whole world of problems still remains open, waiting to be explored. Keywords : Symbolic data analysis, Interval data, Imprecise data, Standardization, Clustering, Discriminant Analysis, Parametric modelling of interval data, Statistical tests for interval data, Skew-Normal distribution, (M)ANOVA, Quantile representation.