New Methods for the Classification of inequally distributed Data:

ABC-plots and computed ABC-analysis

J.Lotsch[1], A. Ultsch[2]

[1]Institute of Clinical Pharmacology, Goethe - University, Frankfurt am Main, Germany,
j.loetsch@em.uni-frankfurt.de

[2] Databionics Research Group, University of Marburg, Marburg, Germany,
ultsch@informatik.uni-marburg.de

The assessment of inequal distributions aiming at the selection of only the relevant items is an important step in data mining of high dimensional data [Foster/Stine 04]. Typical examples are the selection of relevant components for a principal component analysis (PCA) [Jolliffe 02] repectively independent component analyses (ICA) [Hendrikse et al 07], or the selection of variables used in symbolic classifiers in Machine Learning (e.g. CART, ID3 etc.) [Guyon/Elisseeff 03]. Procedures to identify the important few items (e.g. eigenvalues, variables, components) as opposed to the "trivial many" [Pareto 09, Juran 75] rely often on "cookbook recipes". This means the selection is based on heuristics with subjectively chosen and often unreported criteria. Recently ABC-plots and the computed ABC-analysis have been introduced [Ultsch/Lotsch 2015] and published in form of a R library on CRAN. ABC-plots display Lorenz-curves in a way that was already used by Lorenz himself in 1905 [Kleiber 05]. ABC-plots allow a sensible comparison criterion of inequal distributions with a suitable Uniform distribution, rather than with the unrealistic Identity distribution [Coulter 89]. The computed ABC-analysis is an algorithmic parameter-free classification of a distribution into distinct sets of the important versus the unimportant variables. In this work the properties of ABC-curves, the ABC-plot and the computed ABC-analysis will be presented. Applications of these methods to typical distributions found in data mining and knowledge discovery, in particular in "Big Data" from life sciences are given.

Coulter, P.B. Measuring Inequality, Westview Press, 1989.

Foster, D.P., Stine, R.A. Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy, Journal of the American Statistical Association, 99 pp. 303–313, 2004.

Guyon, I., Elisseeff,A. (Eds.) Special Issue on Variable and Feature Selection, Journal of Machine Learning Research, 2003.

Hendrikse, A.J., Veldhuis, R.N.J., Spreeuwers, L.J. Component ordering in independent component analysis based on data power, 28th Symp.Information Theory in the Benelux, pp. 211-218, Enschede, The Netherlands, 2007.

Jolliffe, I.T. Principal Component Analysis, Springer Series in Statistics, 2nd ed., Springer, NY, 2002.

Juran, J.M. The Non-Pareto Principle- Mea Culpa, Quality Progress, 8, pp 8-9, 1975.

Kleiber, C. The Lorenz Curve in Economics and Econometrics, Technical Report, University of Dortmund, No.3, 2005.

Pareto, V. Manuale di economia politica, Milan: Società editrice libraria, revised and translated into French as Manuel d'économie politique. Paris: Giard et Briére, 1909.

Ultsch, A., Lotsch, J. Computed ABC analysis for rational selection of most informative variables in multivariate data, PlosOne, in press, 2015.