New Methods for the Classification of Inequally Distributed Data: ABC-plots and computed ABC-analysis

J.Lotsch¹, A. Ultsch²

¹Institute 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] repsectively 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 [Ulltsch/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.


