Random forests followed by ABC analysis as a feature selection procedure for machine-learning

Jörn Lötsch[1,2] and Alfred Ultsch[3]

[1] Goethe - University, Institute of Clinical Pharmacology, Theodor – Stern - Kai 7, 60590 Frankfurt am Main, Germany

[2] Fraunhofer Institute for Molecular Biology and Applied Ecology IME, Project Group Translational Medicine and Pharmacology TMP, Theodor – Stern - Kai 7, 60596 Frankfurt am Main, Germany

[3] DataBionics Research Group, University of Marburg, Hans – Meerwein – Straße 6, 35032 Marburg, Germany

Data acquisition for biomedical research becomes increasingly complex due to the increasing molecular and clinical knowledge of pathomechanisms of diseases. Data from DNA and other biomolecular measurements are typically high dimensional with variable numbers in the range of 10^2 to 10^6. If an understanding of the biological mechanisms or processes is one of the goals of data analysis, then a rational selection of most informative variables is a necessity. Random forest machine learning employs a multitude of decision trees to learn a highly irregular combination of features [1]. It is usually employed for classifier creation. Common technical implementations such as R-libraries (e.g. [2]) output quantitative measures of the importance of each feature for the overall classification performance. These measures are given as the mean decrease in classification accuracy or in the Gini impurity when the respective variable was excluded from random forest building. An optimal selection of the most informative variables can be achieved by the calculated ABC analysis [3] of the importance measures. ABC analysis is a categorization technique for (positively) skewed distributions. It identifies the most important subset among a larger set of items, aiming at dividing a set of data into three disjoint subsets called “A”, “B” and “C”.

Subset “A” comprises the profitable values, i.e., "the important few" [4], i.e., the features selected for classifier building.

The procedure proved suitable for creating symbolic classifiers from high-dimensional laboratory data that are accessible to topical expert interpretation. For example, a classifier respectively biomarker for multiple sclerosis was created by map the input data space comprising serum concentrations of n = 43 different lipid-markers of various classes to the diagnostic classes of either MS patients (n = 102) or healthy subjects (n = 301). Feature selection using random forests and ABC analysis identified n = 7 biomarkers in ABC set “A” for a biomarker based on Bayesian statistics that provided a classification accuracy of 96 – 97 % in training and test data sets. The combination of random forests followed with ABC analysis for feature selection is to our knowledge a novel nonparametric method suitable as a preliminary step for classifier building by various methods.

Funding

This work has been funded by the Landesoffensive zur Entwicklung wissenschaftlich - ökonomischer Exzellenz (LOEWE), LOEWE-Zentrum für Translationale Medizin und Pharmakologie (JL). The funders had no role in method design, data selection and analysis, decision to publish, or preparation of the manuscript. The authors have declared that no competing interests exist.

References

1. Breiman, L.: Random Forests. Mach. Learn. 45, 5-32 (2001)
2. Liaw, A., Wiener, M.: Classification and Regression by randomForest. R News 2, 18-22 (2002)
3. Ultsch, A., Lötsch, J.: Computed ABC Analysis for Rational Selection of Most Informative Variables in Multivariate Data. PLoS One 10, e0129767 (2015)

4. Juran, J.M.: The non-Pareto principle; Mea culpa. Quality Progress 8, 8-9 (1975)