Stability Selection for Metabolomics Data

Ron Wehrens

Computational Biology
Fondazione Edmund Mach
Italy

Abstract

In the field of metabolomics, one aims to obtain a holistic view of all small molecules (metabolites), hopefully providing information on relevant biological processes in the system under study. Data are typically measured by hyphenated mass-spectrometry based detection, leading to thousands of variables for every sample. In the typical patient-control context, this means that we are performing many tests to find relevant differences between the two classes, with the unavoidable risk of false positives, or, alternatively, very little power.

Stability selection (Meinshausen and Buehlmann, 2010), using the lasso as a primary variable selection method, provides a way to avoid many of the false positives in biomarker identification by repeatedly subsampling the data, and only considering those variables as putative biomarkers that consistently show up as important. In our own work, we have shown that also selecting the largest coefficients in non-sparse regression models such as PLS works well, when combined with the stability selection framework (Wehrens et al. 2011). We support these claims with the analysis of several experimental and simulated data sets.

In particular, the BioMark package for R (Wehrens and Franceschi, 2012), implementing stability selection as well as Higher Criticism thresholding, contains an experimental spike-in data set from the area of metabolomics, which can aid in further algorithm testing and development. From these analyses, it follows that stability selection is a very general and robust framework for variable selection.