BIOMARKER SELECTION IN R: THE BIOMARK PACKAGE

Ron Wehrens and Pietro Franceschi

Biostatistics and Data Management, Fondazione Edmund Mach
San Michele all'Adige (TN)
Italy
http://cri.fmach.it/BDM

The introduction of high-throughput measurement techniques has transformed modern biology, where the phrase "high-throughput" should be seen in the light of the number of variables measured rather than the number of samples: that is, we have many data about few objects. Finding differences between two groups of objects (patients and controls, different varieties of fruits, different food products, …) has become a very important area of research. The classical statistical tests in most cases suffer from many false positives from multiple testing, and correcting for this often leads to low power. Moreover, it is unclear what threshold to choose: in many cases the level of alpha is being adjusted until the number of "significant" variables seems more or less right (for whatever reason: follow-up experiments typically are more time-consuming and one may be forced to settle for the best twenty or thirty variables).

The R package BioMark addresses both the problem of what threshold to choose, and how to select relevant variables by providing Higher Criticism thresholding (Donoho and Jin, 2008) and Stability Selection (Meinshausen and Buehlmann, 2010), respectively. The HC approach is a kind of second-level significance testing: it assesses the distribution of p values from a primary test and compares that to the expected uniform distribution. Assuming that real differences are rare and weak, the method then suggests an appropriate cutoff point, based on the data at hand. Stability selection works by repeatedly subsampling the data, and only considering those variables as putative biomarkers that consistently show up as important.

We have adapted both Higher Criticism and Stability Selection for use with omics data, and provide an easily accessible implementation in our BioMark package, available from CRAN (http://cran.r-project.org). The Higher Criticism approach is extended to also work with methods typically employed in metabolomics and proteomics, such as PLSDA, PCLDA and the VIP measure (Wehrens and Franceschi, 2012). Since no null distribution can be defined beforehand, it is derived using label permutations, and the results show excellent performance for experimental spike-in data (also included in the BioMark package). The same multivariate methods have been combined with stability selection (Wehrens et al., 2011) and again show very good results, testifying of the robustness of the stability selection paradigm.

References:

Donoho D, Jin J (2008). "Higher criticism thresholding: optimal feature selection when useful features are rare and weak." PNAS, 105(39), 14790–14795.
Meinshausen N, Buehlmann P (2010). "Stability selection." J. R. Statist. Soc. B, 72, 417–473. With discussion.
Wehrens R, Franceschi P, Vrhovsek U, Mattivi F (2011). "Stability-based biomarker selection." Anal. Chim. Acta, 705, 15–23.
Wehrens R, Franceschi P (2012). "Thresholding for Biomarker Selection in Multivariate Data using Higher Criticism". Submitted for publication.