Group-regularized logistic elastic net regression: improved omics-based classification

Magnus M. Münch[1,2], Carel F.W. Peeters[1], Aad W. van der Vaart[2], and Mark A. van de Wiel[1,3]

April 1, 2018

1. Department of Epidemiology & Biostatistics, Amsterdam Public Health research institute, VU University Medical Center, PO Box 7057, 1007 MB Amsterdam, The Netherlands
2. Mathematical Institute, Leiden University, Leiden, The Netherlands
3. Department of Mathematics, VU University, Amsterdam, The Netherlands

Abstract

Classification problems are common in omics research. Such problems arise, for example, in the design of diagnostic tests, and the prediction of treatment response. Often external information on the omics features is available. Examples of such information sources are: (a) results on the same genes obtained in a previous study (e.g., p-values), (b) information from a publicly available database that summarizes the prior information on the molecular features involved (e.g., the Cancer Gene Census), (c) omic annotation (e.g., the location of a gene on the chromosome) and (d) previously shown importance of features in related problems. Although such information can rarely be directly included in the statistical analysis, it is often useful and thus has the potential to enhance classification performance.

We propose to include external information by a group-regularized (logistic) elastic net regression algorithm. The groups of features are based on the external information, such that each group of features receives its own penalty parameter. The method makes use of the Bayesian formulation of logistic elastic net regression to estimate both the model and penalty parameters in an approximate empirical-variational Bayes framework. By estimating the group-specific penalty parameters from the data, we avoid a priori (i.e., subjective) specification of feature importance.

Simulation results show that in settings where the grouping of the features is informative, group-specific penalization of the features does indeed enhance classification performance. These findings are confirmed in an application of the method in a cancer omics study.