High-throughput Profiling for Quantitative Phenotypes

Cesare Furlanello (1), Giorgio Guzzetta (1,2), Michele Filosi (2,1), Samantha Riccadonna (1), Roberto Visintainer (1,2), Giuseppe Jurman (1)

(1) MPBA group, Bruno Kessler Foundation (FBK), (2) University of Trento

INTRODUCTION

Quantitative phenotypes emerge everywhere in systems biology and biomedicine, e.g. in complex common diseases where high individual variability makes difficult or impossible to separate cases into distinct categories. More generally, it can be envisioned that the problem of identifying predictive classifiers from genotypic data will evolve into fitting a multidimensional phenotype pattern or a phenotype trajectory, for which more sophisticated predictive tools need to be developed.

We propose a new machine learning pipeline for profiling quantitative phenotypes from high-throughput functional genomics data. Replicating [1], markers and non-overlapping sets of markers the phenotype are set from a data-regularization framework derived from the naive elastic net and refinements [2, 3, 4].

The L1L2 algorithm is embedded in a Data Analysis Protocol (DAP) schema to ensure reproducibility, control of sources of variability, and avoid selection bias. The DAP can be chosen for different applications on the basis of the dataset sample size to trade off selection bias reduction and statistical significance and accuracy.

DATA ANALYSIS PROTOCOLS

Application of L1L2 phenotype fitting are demonstrated for a) GSCAN: genome-wide genotype murine data and physiological traits; and b) AGRE: genome-wide genotype human data and a behavioral indicator [5] in a cohort of Autism Spectrum Disorder patients; c) MILE: microarray expression data and markers of Myelo-dysplastic Syndrome (MDS) and its transformation in acute myeloid leukemia (AML). Data tasks specifications are given in Table A.

DAP example (GSCAN, Fig. B): based on the guidelines of the MAQC-II project [6], given 15 development-validation splits, model selection is based on internal 10-fold Cross Validation (CV) on each development dataset.

A critical hidden setting is the L1L2 optimal parameters (λ, ρ, λ). It is solved by an algebraic stability protocol that works also on marker lists of different lengths [7]. Marker panels are of top candidates, ranked and selected for predictive accuracy and stability. In Fig. B, the optimal par set is in the bottom-left corner of the diagnostic graph. The (λ, ρ, λ) values are used to train the algorithm on the entire development set and evaluate its predictions against the corresponding validation set.

For the AGRE and MILE datasets, a simple 5-fold CV resampled 10 times (10x5 CV) was used because of the reduced number of available samples. For MILE, model was built on samples from one site, and validated on remaining data. See Fig. C for AGRE.

The L1L2 algorithm was implemented with the mlpy Python package and protocols were run on the FBK Kore cluster.

SATURATION PROCEDURE

For all datasets, the pipeline also includes a saturation procedure to recover markers highly correlated to those selected by L1L2. Features systematically and significantly associated to the given trait, termed ‘top-ranked’, are first chosen as those with both mean absolute regression weight above a given threshold and selected on a minimum number of experiments. Then, features whose population profiles are highly correlated with top-ranked features (absolute value of correlation coefficient above a given threshold, termed ‘top-correlated’) are also put into the list of selected features. Fig. D shows that higher correlation is associated to higher regression weights.

Thus, the saturation procedure recovers features highly correlated to the top-ranked and also highly-ranked in terms of regression weight.

This black box algorithm is similar to analysis by Linkage Disequilibrium after one biomarker is found. Saturation is useful to define non-punctual loci of interest on the chromosomes, e.g. in the case of SNP data.

ACKNOWLEDGMENTS


REFERENCES