Stable Feature Selection for Biomarker Discovery: Use of Biological Information

SANAVIA T (1), AIOLLI F (2), DA SAN MARTINO G (2), BISOGNIN A (3), DI CAMILLO B (1)

(1) Department of Information Engineering, University of Padova, Padova
(2) Department of Pure and applied Mathematics, University of Padova, Padova
(3) Department of Biology, University of Padova, Padova

Motivation
Studies on the transcriptome done with microarray data have experienced a huge diffusion and the analysis of these data has been promising for the identification of key genes which can be used as diagnostic/prognostic markers for the disease. In particular, supervised classification analysis has been largely applied to address this issue and the state-of-the-art machine learning methods have demonstrated to give solutions with empirically good accuracy. However, the stability of the selected biomarkers is also a very important task. If an accurate system tends to select the same biomarkers in different independent experiments, then it is more likely that the selected biomarkers are the right ones. In general, there are two stability issues arising in gene expression classification and analysis. First, it would be desirable that classifiers obtained using different training data do not change too much (low variance of the classifiers). Since training data are often limited, predictive models obtained from different datasets can be extremely different. Secondly, since the number of features is generally very high, then features can be combined in many different ways to give solutions able to explain data. As a consequence, many possible sets of features can be considered relevant to the task and equally good in terms of accuracy. This characteristic makes the process of selecting the set of relevant features which are relevant for a classification task a very hard problem. A possible approach to improve list stability is to integrate biological information from genomic databases in the learning process.

Methods
This work investigates the effects on biomarker list stability of the integration in the learning process of different biological information like functional annotations, protein-protein interactions, and expression correlation among genes. Biological information is codified into similarity matrices between features and the feature space is transformed such that the more similar two features are, the more closely they are mapped. Similarity matrices are defined using eight different metrics specific for the type of biological information used: semantic similarities (Pesquita et Al., 2008) for the annotations on
Gene Ontology Molecular Function and Biological Process; four different measurements of topological similarity for protein-protein interactions extracted from Human Protein Reference Database (Prasad et al., 2009), namely the normalized geodesic distance, the Jaccard coefficient, the Functional Similarity (Chua et al., 2006) and the Probabilistic Common Neighborhood Similarity (Cho and Zhang, 2010); pairwise Pearson and Spearman correlation and mutual information for gene expression data. A linear classifier resembling the Bayes Point Machine in (Herbrich et al., 2001) is used as classification tool and the classifier is applied using a bootstrap approach. The vector of weights produced by this algorithm is used to rank the features and obtain a list of biomarkers. The assessment of the results obtained for different similarity matrices is based on the trade-off between predictive accuracy and feature ranking stability, measured using the Canberra distance (Jurman et al., 2008).

Results
The method has been applied on three publicly available breast cancer datasets. All types of biological information are able to decrease the average Canberra distance over the biomarker lists with respect to the standard classification approach across the resampled datasets generated by the bootstrap method. In particular, matrices obtained from semantic similarity measures on GO annotations and from the normalized geodesic distance on protein-protein interactions are the best performers in improving list stability (Canberra distance decreases by 45%) maintaining almost equal prediction accuracy (variations ranging between -4% and +1%), which is preserved at high levels (more than 80% of accuracy for all the types of similarity matrices). These results are confirmed when the three breast cancer datasets are compared. Protein-protein interaction codified by the normalized geodesic distance and the Gene Ontology Molecular Function and Biological Process similarity matrices are confirmed as the best performing kinds of prior knowledge in terms of stability. The performed analysis supports the idea that when some features are strongly correlated to each other, for example because they belong to the same biological process, then they likely have similar importance and are equally relevant for the task at hand. In other words the weight vector obtained for a classification task should have similar values on indices relative to similar genes. Obtained results can be a starting point for additional experiments and to combine similarity matrices in order to obtain even more stable lists of biomarkers.

Contact email
dicamill@dei.unipd.it