Robust biomarker identification for cancer diagnosis using ensemble feature selection methods

Thomas Abeel[1,2], Thibault Helleputte[3,4], Yves Van de Peer[1,2], Pierre Dupont[3,4], and Yvan Saeys[1,2]

[1]Department of Plant Systems Biology, VIB, Technologiepark 927, 9052 Gent, Belgium.
[2]Department of Molecular Genetics, Ghent University, Gent, Belgium.
[3]Department of Computing Science and Engineering INGI, Université catholique de Louvain, Belgium.
[4]Machine Learning Group, Université catholique de Louvain, Belgium.
{thomas.abeel|yves.vandepeer|yvan.saeys}@psb.ugent.be
{thibault.hellepute|pierre.dupont}@uclouvain.be

1 Introduction

Biomarker discovery is an important topic in biomedical applications of computational biology, including applications such as gene and SNP selection from high dimensional data. Surprisingly, the stability with respect to sampling variation or robustness of such selection processes has received attention only recently. However, robustness of biomarkers is an important issue, as it may greatly influence subsequent biological validations. In addition, a more robust set of markers may strengthen the confidence of an expert in the results of a selection method.

2 Methodology

Our first contribution is a general framework for the analysis of the robustness of a biomarker selection algorithm. Secondly, we conducted a large scale analysis of the recently introduced concept of ensemble feature selection, where multiple feature selections are combined in order to increase the robustness of the final set of selected features. We focus on selection methods that are embedded in the estimation of support vector machines (SVMs). SVMs are powerful classification models that have shown state-of-the-art performance on several diagnosis and prognosis tasks on biological data. Their feature selection extensions also offered good results for gene selection tasks.

3 Results

We show that the robustness of SVMs for biomarker discovery can be substantially increased by using ensemble feature selection techniques, while keeping the same classification performances. The proposed methodology is evaluated on four microarray data sets showing increases of up to 27% in robustness of the selected biomarkers. The stability gain obtained with ensemble methods is particularly noticeable for small signature sizes (a few tens of genes), which is most relevant for the design of a diagnosis or prognosis model from a gene signature.