FERAL: Yet Another Network Based Breast Cancer Outcome Predictor

Amin Allahyar and Jeroen de Ridder
Delft Bioinformatics Lab, Delft University of Technology
Mekelweg 4, Delft, The Netherlands
A.Allahyar@tudelft.nl, J.deRidder@tudelft.nl

Abstract

To improve performance in breast cancer outcome prediction, Network based Outcome Prediction methods have been proposed which construct meta-genes (i.e., features) by averaging expression of nearby genes according to given biological network (e.g., KEGG). Using sparse group lasso, we propose FERAL that uses multiple integration operators to summarize genes into meta-genes while simultaneously determine most relevant operators for predicting patient’s outcome. Extensive evaluation revealed that FERAL is markedly more predictive compared to existing NOPs.

1. Introduction

Metastases at distant sites (e.g. in bone or lung) is the major cause of death in breast cancer patients [5]. According to hallmarks of cancer, deregulation of several processes or cellular pathways is responsible for development of this disease [2]. This premise motivated researchers to aggregate functionally related genes in order to produce meta-genes (i.e., features) with added discriminative power and biological relevance. Relations between genes are often determined by a pre-defined biological network. Hence, these approaches are often called Network based Outcome Prediction (NOP) methods [4].

Contrary to previous claims, recent studies reported that many NOPs do not outperform a model trained over single gene features [4]. This is certainly contradicts with principle of hallmarks of cancer. The main goal of this paper is to identify and alleviate several fundamental issues in current NOPs. For example, we find that the main bottleneck in current NOPs is a poor choice of average operator for integrating the expression of related genes. Moreover, single operator may not be sufficient to capture the aberration of higher level functions in cell. Finally, we noticed that decoupling the training of the classifier from selection of genes/meta-genes hamper the.

We propose FERAL (DelFt nEtwoRk bAsed cLassifier) [1] that exploits Sparse Group Lasso (SGL) [3] to achieve simultaneous selection and integration of genes and meta-genes (see Figure 1). Furthermore, instead of being limited to average, FERAL exploits a wide range of operators including a previously unexplored supervised integration strategy. Finally, we present the results of extensive experiments in which FERAL achieved statistically significant performance improvement along with relevance to known cancer gene sets. Taken together, these feats enable biological interpretation of the trained classifier which, results in highly relevant mechanistic insights.

2. Materials & Methods

As discussed, most NOPs employ the average operator for constructing their meta-genes [4]. However, if enclosing genes exhibit opposite association w.r.t. the class label, this operator can even cancel out their individual predictive contribution. To alleviate this issue, we propose the Direction Aware Average (DA2) operator which adjusts the direction of genes before taking the average. DA2 is defined as:

\[
DA2_g = \frac{1}{|\Psi_g|} \sum_{j \in \Psi_g} \text{sgn}(C_j) \times E_j,
\]

where \( \Psi_g \) is the gene set of seed gene \( g \) and \( E_j \) and \( C_j \) contain the expression and correlation values with the class label of gene \( j \), respectively. We also included other biologically inspired operations such as the max/min (to model
AND/OR relations) or the variance (to capture variability of expression levels) (see Figure 2). FERAL leverages the SGL to select the most relevant features in both gene and meta-gene levels. Each group of SGL encloses a gene and nine of its closest neighbors (according to given network) as well as corresponding meta-genes.

3. Results and discussion

For evaluation, we used the Amsterdam Classification Evaluation Suite (ACES) [4], a cohort of 12 studies in NCBIs Gene Expression Omnibus (GEO). The utilized networks include: I2D [4], co-expression and random network.

3.1. Performance comparison

Figure 3 shows the obtained average AUCs for 10 repeats of the subtype stratified cross-validation. Even though existing methods can easily be improved by including a simultaneous selection and integration step, we observe that FERAL still offers superior performance across all three networks considered. This performance improvement is very significant (p-value < 7 · 10⁻⁸; paired t-test with the best other method iPrk using the co-expression network). This demonstrates that, on top of the SGL approach, it is beneficial to provide the classifier with a rich collection of meta-genes based on different aggregation strategies.

3.2. Functional enrichment of marker genes

To assess whether the methods under study are capable of detecting relevant markers, we evaluate the concordance of sets of known cancer-related genes with the ranked set of genes produced by each method using the AUC measure. The observed enrichments obtained using the I2D network is demonstrated in Figure 4. This experiment show that all methods have very modest enrichments. The notable exception is FERAL, which is vastly superior and close to 0.7 for most cancer-related gene sets. Taken together, these observations support those made in Section 3.1, that is, incorporating network information does not greatly improve performance, but it does contribute to stabilizing the marker gene sets and finding the biologically relevant genes.

References