ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Discriminant functional gene groups identification with machine learning and prior knowledge

Grzegorz Zycinski[1] and Margherita Squillario[1] and Annalisa Barla[1]
and Tiziana Sanavia[2] and Alessandro Verri[1] and Barbara Di Camillo[2]

1- Department of Computer and Information Science
University of Genoa
via Dodecaneso 35, I-16146 Genova, Italy

2- Information Engineering Department
University of Padova
via Gradenigo 6A, I-35131 Padova, Italy

Abstract. In computational biology, the analysis of high-throughput data poses several issues on the reliability, reproducibility and interpretability of the results. It has been suggested that one reason for these inconsistencies may be that in complex diseases, such as cancer, multiple genes belonging to one or more physiological pathways are associated with the outcomes. Thus, a possible approach to improve list interpretability is to integrate biological information from genomic databases in the learning process. Here we propose KDVS, a machine learning based pipeline that incorporates domain biological knowledge a priori to structure the data matrix before the feature selection and classification phases. The pipeline is completed by a final step of semantic clustering and visualization. The clustering phase provides further interpretability of the results, allowing the identification of their biological meaning. To prove the efficacy of this procedure we analyzed a public dataset on prostate cancer.

1 Background

In the last decade, transcriptome analysis performed with high-throughput microarrays has experienced a huge diffusion in disease classification, where in a typical experimental design data come from different subjects and phenotypes. In this context, classification methods are often used to select biomarker genes useful for answering diagnostic, prognostic and functional questions related to a disease [1]. However, high-throughput analysis carried out in different laboratories and research centers has given different results, with limited overlap or reduced statistical significance [2, 3]. These differences are matters of important scientific discussions and are imputed to two main reasons: *(1)* datasets often include small numbers of subjects (some tens) with respect to the number of variables (tens of thousands of probes for human genome); *(2)* the most complex pathologies, such as cancer, are heterogeneous and multi-factorial, as a result of the alteration of multiple regulatory pathways, rather than referable to a single dysfunctional gene like in monogenic diseases [4]. As a consequence, data are characterized by many correlated features, which lead to different but

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

equivalent solutions of the classification/feature selection task. The low reproducibility of biomarker lists strongly affects the biological interpretability of the results. To address this issue, the standard approach is *enrichment analysis*, which uses domain knowledge (*e.g.*, annotations from Gene Ontology - GO or pathways from the Kyoto Encyclopedia of Genes and Genomes - KEGG) to retain only those functional groups of genes significantly represented in the signature [5]. This approach allows for a better interpretability, but it tends to promote only those functional groups that include a high fraction of the selected genes in the signature. In the present work we propose a method to select the most discriminant functional groups of genes based on $\ell_1\ell_2$ regularization with double optimization, as described in [6]. The current implementation (Knowledge Driven Variable Selection, KDVS) uses the GO graph [7] as prior knowledge to be injected in the variable selection procedure. For each functional group of genes annotated in GO, KDVS extracts its corresponding data matrix from the available measures (*e.g.*, expression data from microarray) and applies an $\ell_1\ell_2$ classification/feature selection step, estimating a prediction error and a list of relevant variables in the group and associating the input data (expression) to the output data (phenotype) with a classification model. Splitting the classification task into sub-problems on sub-groups allows addressing issue *(1)*, *i.e.*, data dimension, whereas $\ell_1\ell_2$ regularization allows accounting for correlation among genes, thus addressing issue *(2)*. The KDVS output is a set of relevant GO terms, which describe the most active molecular functions or biological processes correlated to the phenotype. The case study analyzed in this work is a public microarray dataset on Prostate cancer.

2 KDVS: prior knowledge and feature selection

Data and Prior Knowledge integration framework. The general schema of KDVS is presented in Figure 1. It is based on the prototype presented in [8], implemented in Python. It includes a raw data processing framework for normalization and summarization, using the state-of-the-art algorithms for high-throughput microarray technologies [9] and a *local integration framework* to integrate microarray platform annotations [10] with prior biological knowledge from GO. The schema in Figure 1 is flexible enough to incorporate different data (Array Express - AE, Gene Expression Omnibus - GEO) and knowledge sources (GO, KEGG). In the following, we refer to GO source. The result of this phase is a dynamically created *information ensemble*, implemented as mashup of relational database and file objects. For every GO term, we collect the set of corresponding probesets, and extract expression values across all samples considered for the classification/regression task. Therefore the original $n \times p$ data matrix X_{tot} (n samples, p variables) is partitioned in submatrices X, one per GO term. The submatrices are stored as well as in the *information ensemble*.

Statistical framework. The following step, *i.e.*, *the statistical framework*, is divided in two steps: $\ell_1\ell_2$ feature selection and semantic clustering.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

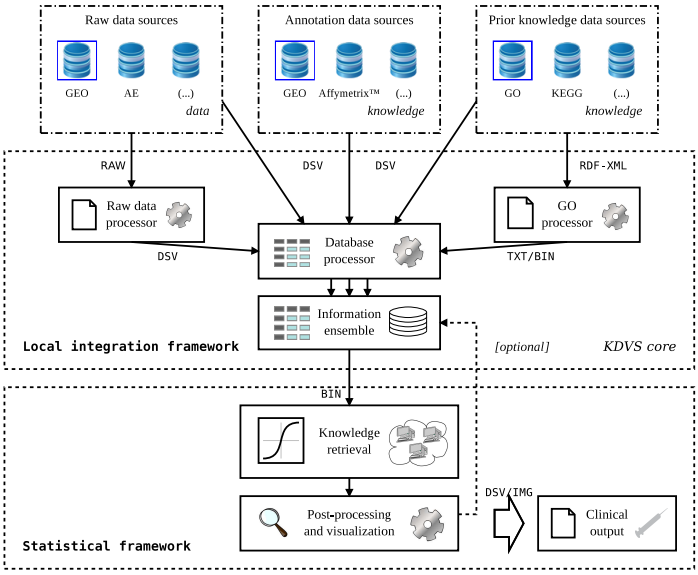

Fig. 1: General schema of Knowledge Driven Variable Selection prototype system. Highlighted sources were used in the current implementation.

$\ell_1\ell_{2\,FS}$ feature selection. The method is based on the optimization principle presented in [11] and it looks for a linear function (model), whose sign gives the rule that can be used to associate a new sample to its predicted phenotype. For each submatrix X and corresponding phenotype vector Y, the algorithm finds the β that minimizes: $\|Y - X\beta\|_2^2 + \tau \|\beta\|_1 + \mu \|\beta\|_2^2$, using an *iterative soft thresholding* algorithm. The output function is a sparse model, *i.e.*, some input variables will not contribute to the final estimator evaluated by regularized least squares (RLS) only on the relevant features, as in [6]. The least square term ensures fitting of the data whereas adding the two penalties allows avoiding overfitting. The role of the two penalties is different: the ℓ_1 term enforces the solution to be sparse, the ℓ_2 term preserves correlation among the variables. The training for selection and classification requires the choice of the regularization parameters for both $\ell_1\ell_2$ regularization and RLS denoted with τ^* and λ^*, respectively. In fact model selection and statistical significance is performed within two nested K-cross validation loops as in [12]. If the estimated error is below a fixed threshold, the submatrix X and the corresponding GO term are selected as meaningful. In addition, for each submatrix, we obtain a list of selected genes. The final output from *knowledge retrieval* is the list of *selected* GO terms, their estimated prediction error, and the list of relevant genes within each term.

Semantic clustering. To increase interpretability of the results, we apply a hierarchical agglomerative clustering (average linkage) to selected GO terms. We have chosen the average linkage method because it allows avoiding nested clus-

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

ters or cluster which can lead to highly inhomogeneous biological information, as it could happen in complete and single linkage. However, the method allows the user to choose the preferred linkage method to use in the semantic clustering. We used the Resnik semantic similarity [13] normalized to the maximum observed value to assess the degree of relatedness between two GO terms c_1 and c_2: $Sim_{Resnik}(c_1, c_2) = max_{c \in MICA(c_1,c_2)} IC(c)$, where IC is the Information Content [14] and $MICA$ indicates the most informative common ancestors of terms c_1 and c_2 in the GO directed acyclic graph (DAG). The IC for the GO term c is defined as: $IC(c) = -log\left(\frac{freq(c)}{freq(root)}\right)$ *i.e.*, the negative logarithm of the ratio between the frequency of the term t in a corpus of annotations (i.e., the number of times the term t and each of its descendants occur in GO annotation) and the frequency of the root term (corresponding to the sum of the frequencies of all GO terms). The final output from this post-processing step is a set of semantically homogenous clusters of GO terms. Since Resnik similarity measure between two GO terms is based on the information content of the common ancestor, semantic similarities of generic and few informative terms are low and these terms are clustered together with their descendent GO terms in the GO graph. In this way, the semantic clustering performed using this measure avoids the creation of a cluster formed by only generic GO terms with different biological meaning, which can occur instead when other semantic measures are used, such as Lin's semantic measure which normalizes this similarity with respect to the information content of the two compared GO terms [15].

3 Experimental Results

We analyzed a publicly available dataset (GSE6919, GEO) of prostate tissues measured on the Affymetrix HG-U95Av2 microarray platform. We preprocessed and performed quality control with R package scripts based on the *aroma* package and the *arrayQualityMetrics* library, discarding one sample and finally considering 25 metastatic and 64 primary tumor samples. We applied KDVS, injecting prior knowledge from Molecular Function (MF) GO domain. We considered as significant those GO terms whose balanced classification error is below 30%, thus selecting 167 discriminant GO terms, which were grouped into 12 different clusters of GO terms according to their similar functional meaning by applying a semantic clustering (see Figure 2). Clusters 1, 2 and 3 (see legend in Figure 2) include GO terms involved in Binding of molecules to cell receptors (*e.g., calmodulin, growth factor, insulin-like growth factor binding*) or to other molecules (*e.g., nucleic acid, SH3 domain, NF-kappaB, and transcription factor binding*). These results underline that the functions occurring in tumor cells are related to the binding of key molecules as calmodulin (that mediates several processes as metabolism, inflammation, intracellular movements, immune response), several growth factors (as the insulin-like growth factor binding that has been correlated with the risk of prostate cancer in a large longitudinal study [16]) and other key molecules (involved in signaling pathways regulating the cytoskeleton, the Ras protein and the Src kinase). Interestingly, in Figure 2 these

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

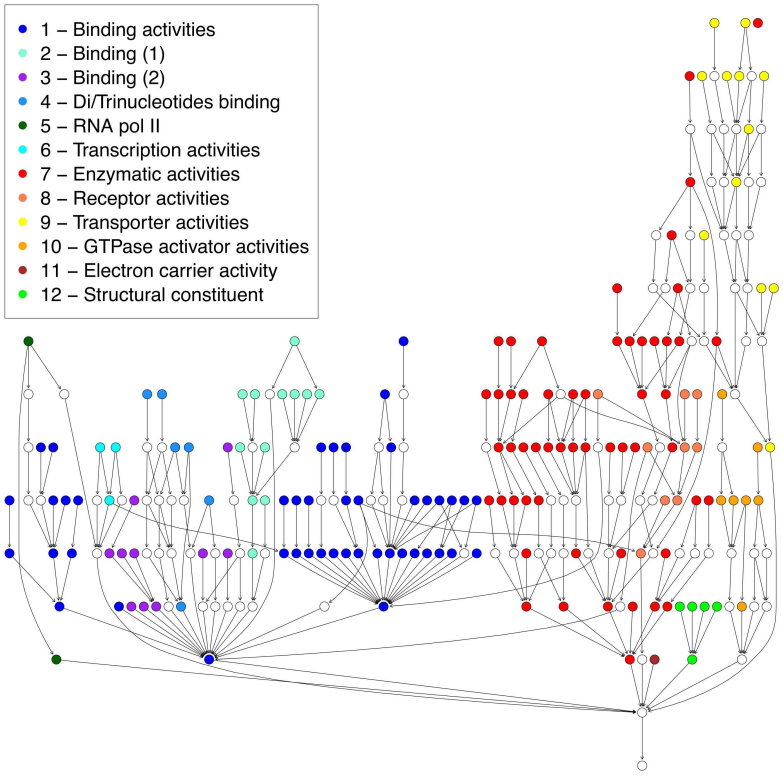

Fig. 2: GO subgraph with clusters identified with semantic clustering in the list of MF nodes for tumor vs. metastases experiment.

clusters are close to each other because of their semantic meaning. The other semantically related clusters are: Di/tri Nucleotide Binding, RNA pol II and Transcription Activities (clusters 4, 5 and 6). The biggest cluster is Enzymatic Activities, including GO terms as *Methyltransferase Activity, ATPase activity, Oxidoreductase Activity, Metalloendopeptidase Activity*). The majority of these enzymes are related to the metabolism. Moreover, metalloproteinases are known to be fundamental for tumor invasion [16]. In the same subgraph (right side of Figure 2) we observe the Transporter Activities cluster (*e.g., calcium channel activity, lipid transporter activity*), the Structural Constituent (*e.g., structural constituent of cytoskeleton, extracellular matrix structural constituent*) and GTPase activator activity (including few other enzymes *e.g., Rho guanyl-nucleotide exchange factor activity, small GTPase regulator activity*). The involvement of several enzymatic classes suggests that all of them are fundamental to meet the needs of the aggressive tumor cells.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

4 Conclusions

When comparing classes of subjects belonging to different phenotypes using genomic data, biological annotation of features can give immediate and intuitive information of the phenomenon under investigation. KDVS provides a method able to integrate biological prior knowledge (GO in this context) into statistical class prediction analysis of high-throughput data. KDVS gives, at a glance, a direct overview of the relevant functions and processes characterizing the biological problem under study, avoiding post-processing functional analyses which can alter the correct biological interpretation of the results.

References

[1] I Guyon and A Elisseeff. An introduction to variable and feature selection. *J Mach Learn Res*, 3:1157–1182, 2003.

[2] L Ein-Dor, I Kela, and G Getz et al. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171–178, 2005.

[3] AL Boulesteix and M Slawski. Stability and aggregation of ranked gene lists. *Briefings in Bioinformatics*, 10(5):556–568, 2009.

[4] X Solé, N Bonifaci, and N López-Bigas et al. Biological Convergence of Cancer Signatures. *PLoS ONE*, 4(2):e4544, 02 2009.

[5] J.-H Hung, T.-H Yang, and Z Hu et al. Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in Bioinformatics*, Sep 2011.

[6] C De Mol, S Mosci, and M Traskine et al. A Regularized Method for Selecting Nested Groups of Relevant Genes from Microarray Data. *J. of Comp. Bio.*, 16:1–15, Apr 2009.

[7] M Ashburner, C A Ball, and J A Blake et al. Gene Ontology: tool for the unification of biology. *Nat Gen*, Jan 2000.

[8] G Zycinski, A Barla, and A Verri. SVS: Data and knowledge integration in computational biology. *Proc. of IEEE EMBC 2011*, 2011.

[9] R C Gentleman, V J Carey, and D M Bates et al. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004.

[10] R Edgar, M Domrachev, and A E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Res*, 30, 2002.

[11] H Zou and T Hastie. Regularization and variable selection via the elastic net. *J Roy Stat Soc B*, Jan 2005.

[12] A Barla, S Mosci, and L Rosasco et al. A method for robust variable selection with significance assessment. *Proceedings of ESANN 2008*, Jan 2008.

[13] P Resnik. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *J Artif Intell Res*, 11, 1999.

[14] P W Lord, R D Stevens, and A Brass et al. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–83, Jul 2003.

[15] Lin D. An information-theoretic definition of similarity. *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304, 1998.

[16] W A Schulz. *Molecular Biology of human cancers.* Springer, 2007.