Towards Formal Interpretation of Linear Models Learnt from Genome-wide Data

Michael ANDĚL[1]

[1]Dept. of Computer Science, Faculty of Electrical Engineering, Czech Technical University, Technická 2, 166 27 Praha, Czech Republic

andelmi2@fel.cvut.cz

Abstract. *Current high-throughput technologies lead to the boost of omics data with thousands of features measured in parallel. The phenotype specific markers are learned from the data to better understand the disease mechanism and to build predictive models. However, the learning is prone to overfitting, caused by a small sample size and large feature space dimension. Consequently, resulting models are inaccurate and difficult to interpret due to the complex nature of omics processes. In this paper we present the discoveries and nuggets we have made by tuning regularization parameters of our method we have recently developed. We extracted nuggets supported by relevant literature record. The main contribution is that these nuggets are relevant to potentially causal mutations, though extracted from solely gene expression, i.e. non-mutational data.*

Keywords

support vector machine, gene expression, machine learning.

1. Introduction

Life sciences, namely molecular biology, have undergone an explosive progress during recent years. Along with the human genome decoding and subsequent conjuncture of high-throughput technologies, such as *microarrays* or next-generation sequencing, a mass production of biomedical data has begun. The high-throughput technologies facilitate parallel measurement of thousands of features at the multiple *omics* levels such as genomics, transcriptomics, proteomics or epigenomics. Consequently, thanks to the still-improving knowledge of omics units structure, we are able to algorithmically infer underlying regulatory mechanisms and omics interactions. These predicted interactions and mechanisms are available along with the databases of in vitro validated interactions and mechanisms, growing solely due to conventional scientific activity.

However, to really profit from this progress, namely to deliver personalized medicine [17] in the field of healthcare and facilitate new discoveries in the research field, respectively, there is an urgent need for well-defined and reproducible methodologies of evaluating and interpreting the masses of data generated and driving the experiments in the smart way. Currently, we are flooded with diverse results of various studies and respective methods of evaluation. The results are often false positive, due to the overfitting on relatively small sample of examples. Moreover, they often lack any mechanistic or even causal interpretation. A modern evaluation technique should not only assign significance to predefined hypothesis, but should be capable of 1) active making new hypotheses based on *measured* data and available *domain knowledge*, 2) estimating their empirical validity and 3) proactive spotting new targets for their causal verification in vitro.

In our latest work, we have developed a methodology for learning comprehensible yet accurate linear models [2]. The models are based on support vector machine (SVM), while their comprehensibility and interpretability is delivered by several regularization terms. In this paper we present the discoveries and nuggets we have made by tuning regularization parameters of our method.

2. Motivation and Related Work

Here, we give a motivation for using domain knowledge in omics data analysis. Next, we present regularization as a promising approach to integrate the knowledge directly into the learning in order to increase the classification accuracy and comprehensibility of resulting molecular markers. Then, we review the existing regularization methods for linear models.

2.1. Integration of Domain Knowledge

The straightforward approach for integration of domain knowledge is *set-level analysis* [16, 14, 9], where one aims to identify entire sets of genes that are differentially expressed. The sets typically refer to a priori defined pathways or gene-ontology (GO) terms. The expression level of contained

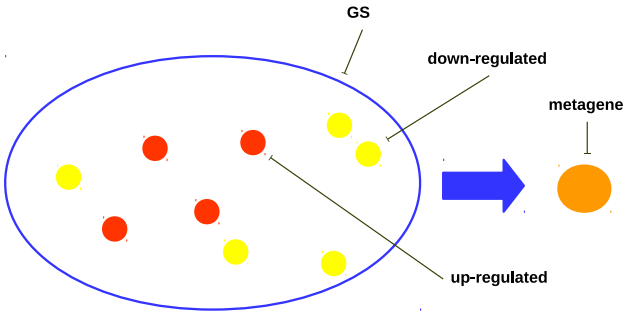

Fig. 1. Undesired effect of aggregation, when not the majority of genes in a gene set is expressed.

genes is aggregated to one meta-feature. These new features are obviously more interpretable as they refer to previously defined abstract terms. Actually, this is a dimension reduction approach to prevent the overfitting. Surprisingly, the overall predictive accuracy often decreases [13]. Though removing noise, a portion of predictive information may be erased by an aggregative procedure. This phenomenon, depicted in Figure 2.1, may appear namely in the case when a significant number of genes is expressed in a predefined gene set, yet not the majority. To address this a more granulated gene sets should be employed, e.g. the more specific GO terms. However, the number of possible sets in which the genes aggregate can be large too (more than 10,000 curated gene sets currently exist).

An alternative approach to incorporate domain knowledge is *regularization*. Regularization is a general way to address overfitting by imposing a prior hypothesis on the learning process. The *hypothesis* may be interpreted simply as a restriction of the model space as in the case of decision tree pruning. For our interest, the hypothesis has a form of domain knowledge. Unlike set-level aggregation which crisply changes and reduces the set of available models, the regularization approach merely gives a soft preference for one model over the less expected one. The latter may be admitted though, requiring stronger evidence, namely more training examples in its favour.

The regularization approach allows a more flexible format of domain knowledge than the set-level aggregation. Unlike e.g., predefined pathways which are purely synthetic human-made concepts, the regularization-based framework enables us to work with more natural entities such as the omics feature interactions. The interactions are commonly based on structural properties of the units that underlie the features, such as nucle- otide sequence or a higher-order protein structure. Yet, it does not say which data context (e.g., organism, phenotype, disease) the interaction works in. This issue of contextual dependency is addressed by the regularization approach as follows. During the learning process, if a feature empirically proves important for the phenotype prediction, we impose that some of the *potentially* interacting features would be similarly important. Finally, if a pair of interacting features proves relevant in the same phenotype model, we can assert this interaction is valid within the context of particular domain and phenotype. Even though the predicted interactions suffer from false positivity, the method needs not be misconducted. The regularization, as introduced above, does not insist on all the interactions, unless they are sufficiently supported by the training examples.

2.2. Regularization of Linear Models

Regularization is mainly exploited in statistical learning. Support vector machine (SVM) [6] is a state-of-the-art statistical learning method which performs well even in GE domain, being capable of dealing with large dimensionality with sufficient generalization [8]. Nevertheless, in life sciences, the model itself is often as highly appreciated as its predictive output. The support vector machines provide a black-box model, though, which is hard to be explained reasonably by domain experts. Hence, there have been several attempts to brighten SVM-based black box models by additional regularization terms. Similarly to *elastic-net* regularization [20] in the case of regression, the *doubly regularized SVM* [18] adds a *sparsity term* to the objective function. This term imposes ℓ^1-norm to the weights of linear model, and thus encourages some of them to be zero. This leads to potentially more interpretable models. Actually, the learner performs an embedded *feature selection* (FS) without common feature-selection bias [1].

When regularizing by the domain knowledge, the straightforward approach is using *network kernels*. Namely, Lavi [10] employs a special *interaction term* to regularize SVM. This term encourages an optimizer to assign similar weights to the interacting features consistently with the intuitive definition of knowledge-based regularization in Section 2.1.

A similar approach is in [5]. However, those SVM-based methods provide the models which are still difficult to interpret. The resulting model is merely a set of weights for *all* the measured features, some higher, some lower. To obtain a comprehensible model, feature selection is obviously needed. Anyway, as mentioned above, FS introduces another kind of bias into the model. Moreover, these methods have been reported as being quite unstable as to the network strength parameter β [10], while there is absolutely no rule of thumb to set this parameter in advance. Hyperparameters like this are often set by nested cross-validation which increases the computational cost quadratically, though.

Li [11] nicely combines the interaction and sparsity term in the linear regression analysis. More interpretable models are thus produced. This problem is reducible to

the lasso problem [15] which can be effectively resolved. However, as a regression formulation, this approach is not suitable for classification problems. Moreover, there is no study of model behaviour under different parametrizations. Henceforth, we present a general methodology for learning the interpretable interaction-based classification models, with the regularization parameters set by the expert's intuition. The embedded feature selection enables prospective identification of key processes related to the phenotype.

3. Materials and Methods

In this section, we briefly summarize the methodology we had previously developed. Then, we describe the parameter-tuning process we used to yield the discoveries.

3.1. Sparse Network-regularized SVM

We employed the approach called Sparse Network-regularized SVM (SNSVM) we had previously developed [2]. Our approach is similar to the network-constrained linear regression [11], but designed for the classification tasks. By merging all the regularization terms described in Section 2.2 we define the SNSVM as:

$$\min_{w_0,w_1,\ldots,w_M} \sum_{i=1}^{M} w_i^2 + \lambda \sum_{i=1}^{M} |w_i| + \beta \sum_{ij\in\mathcal{I}} A_{ij}(w_i - w_j)^2, \tag{1}$$

$$s.t. : (\mathbf{w}^T \mathbf{x}_i + w_0) y_i \geq 1, \forall i = 1, \ldots, N, \tag{2}$$

where Constraints 2 ensure all the N data points are separated in the space of M features. The first term in Objective 1 is the margin maximization term from the standard SVM. The second is the sparsity term, discussed in Section 2.2, as used in [20, 18, 11]. The third term is the interaction term (see Section 2.2) used by [10, 11, 5]. The constants λ and β, respectively, define the ratio of sparsity and interconnection of features employed in the model. The intention is to keep only a *few* (sparsity term) *meaningfully related* (interaction term) features with *reasonable generalization* (margin term).

3.2. Model Selection

By increasing λ, the model gets sparser as to the number of nonzero-weight features. Whereas increasing β, the model grows larger in a *meaningful* way, i.e., the model expansion follows the network topology. E.g., the evolutionary related genes may be targeted by a common regulatory process. Similarly, the genes with interacting proteins are often co-expressed. If a group of features has already shown a nonzero-weight in the model, it may be viewed as afflicted by the unseen *phenotype-causing* mechanism.

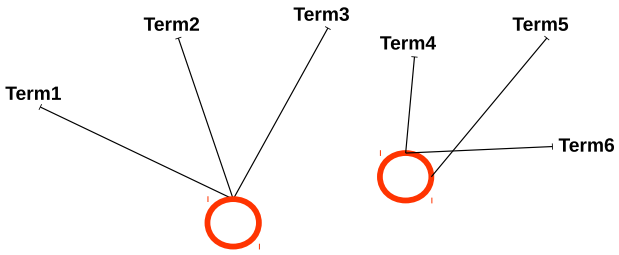

Fig. 2. An example of ultimately sparse model.

Behold an ultimately sparse model learned under the configuration with high value of λ (sparsity) but low value of β (network connectivity). This model, though accurately describing phenotype, is hardly possible to identify with some biological process (see Figure 3.2). Each of the genes, proved with nonzero-weights in resulting model may be annotated by several terms referring to distinct biological processes. However, by relaxing the sparsity through increasing the value of β we can expand the model in a *meaningful* way. The meaningful means in a predefined biological relation such as the protein interactions in our case. The expected result is that the nonzero-weight genes share significant number of annotations with only a few of terms.

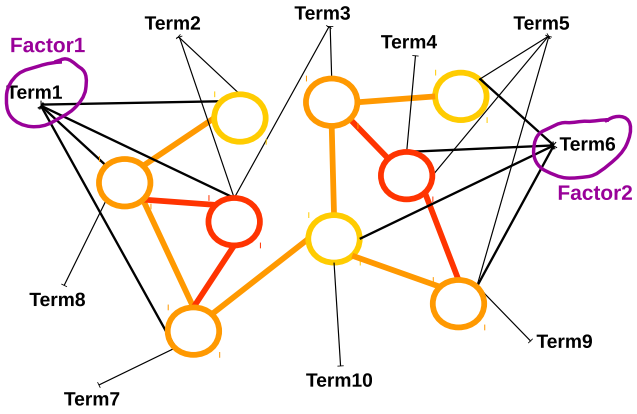

Fig. 3. An example of meaningfully expanded model.

4. Results

By employing the methodology described in Section 3.2 on data related to myelodysplastic syndrome [2] (MDS), we obtained a model (Figure 5) significantly enriched in four GO terms. The term `RNA splic trans ester...` is related to the alternative RNA splicing. The term `Ribonucl compl sun org` refers to the ribonucleotid complex subunit organization. These two terms can be together explained as relevant to the spliceosomal mutations in ribosomes. These mutations occur in approximately half of all MDS patients and seem to be highly specific for this disorder [4].

The next interesting enriched term is relevant to the viral process. Surprisingly a hypothesis of MDS triggered by a virus has been postulated independently once upon a time [12]. This may be seen as a potentially new discovery in the domain.

The last term is related to mitochondrial translation (`Mit transl`). Even this term has a support in literature as independently supporting a potentially novel discovery [19, 7], and may be related to the mutations in mitochondrial DNA.

5. Conclusion

The main contribution is that these nuggets are relevant to potentially causal mutations, though extracted from solely gene expression, i.e. non-mutational data. It suggest that proper integrating the available data with relevant general knowledge may unveil nontrivial relations. Certain of these discoveries are worth for further wet-lab research.

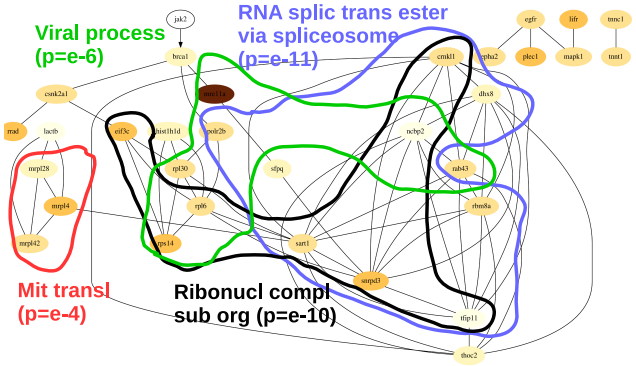

Fig. 4. Resulting model

Acknowledgements

The research was supervised by Jiří Kléma. The work was supported by the grant SGS14/197/OHK3/3T/13 of the Grant Agency of the Czech Technical University in Prague.

References

[1] AMBROISE, C., AND MCLACHLAN, G. J. Selection bias in gene extraction on the basis of microarray gene-expression data. 6562–6566.

[2] ANDĚL, M., MASRI, F., KLEMA, J., KREJCIK, Z., AND BELICKOVA, M. Sparse omics-network regularization to increase interpretability and performance of linear classification models. In *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on* (2015), IEEE, pp. 615–620.

[3] BELLAZZI, R., AND ZUPAN, B. Towards knowledge-based gene expression data mining. *J. Biomed. Inform. 40*, 6 (2007), 787–802.

[4] BOULTWOOD, J., DOLATSHAD, H., VARANASI, S. S., ET AL. The role of splicing factor mutations in the pathogenesis of the myelodysplastic syndromes. 153–161.

[5] CHEN, L., XUAN, J., RIGGINS, R. B., ET AL. Identifying cancer biomarkers by network-constrained support vector machines. 161.

[6] CORTES, C., AND VAPNIK, V. Support-vector networks. 273–297.

[7] GUPTA, M., MADKAIKAR, M., RAO, V. B., ET AL. Mitochondrial DNA variations in myelodysplastic syndrome. 871–876.

[8] GUYON, I., WESTON, J., BARNHILL, S., AND VAPNIK, V. Gene selection for cancer classification using support vector machines. 389–422.

[9] HOLEC, M., KLÉMA, J., ŽELEZNÝ, F., AND TOLAR, J. Comparative evaluation of set-level techniques in predictive classification of gene expression samples. S15.

[10] LAVI, O., DROR, G., AND SHAMIR, R. Network-induced classification kernels for gene expression profile analysis. 694–709.

[11] LI, C., AND LI, H. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics 24*, 9 (2008), 1175–1182.

[12] RAZA, A. Hypothesis: myelodysplastic syndromes may have a viral etiology. 245–256.

[13] STAIGER, C., CADOT, S., GYÖRFFY, B., ET AL. Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis.

[14] SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., ET AL. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. 15545–15550.

[15] TIBSHIRANI, R. Regression shrinkage and selection via the lasso. 267–288.

[16] TUSHER, V. G., TIBSHIRANI, R., AND CHU, G. Significance analysis of microarrays applied to the ionizing radiation response. 5116–5121.

[17] VAN'T VEER, L. J., AND BERNARDS, R. Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature 452*, 7187 (2008), 564–570.

[18] WANG, L., ZHU, J., AND ZOU, H. The doubly regularized support vector machine. *Statistica Sinica 16*, 2 (2006), 589.

[19] WULFERT, M., KÜPPER, A. C., TAPPRICH, C., ET AL. Analysis of mitochondrial DNA in 104 patients with myelodysplastic syndromes. 577–586.

[20] ZOU, H., AND HASTIE, T. Regularization and variable selection via the elastic net. 301–320.

About Authors...

Michael ANDĚL has received BSc degree in Biomedical and clinical technology and MSc degree in Artificial intelligence, both at Czech Technical University in Prague (CTU). Currently, he is a researcher and PhD student in Artificial intelligence and Biocybernetics at CTU, Department of Computer Science and Engineering. His specialization is machine learning from heterogeneous omics data.