Methods 83 (2015) 88–97

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

# Network-constrained forest for regularized classification of omics data

Michael Anděl [a], Jiří Kléma [a,*], Zdeněk Krejčík [b]

[a] Department of Computer Science, Czech Technical University, Technická 2, Prague, Czech Republic
[b] Department of Molecular Genetics, Institute of Hematology and Blood Transfusion, U Nemocnice 1, Prague, Czech Republic

ARTICLE INFO

*Article history:*
Received 30 January 2015
Accepted 2 April 2015
Available online 11 April 2015

*MSC:*
00-01
99-00

*Keywords:*
Omics data
microRNA
Machine learning
Random forest
Domain knowledge
Regularization

ABSTRACT

Contemporary molecular biology deals with wide and heterogeneous sets of measurements to model and understand underlying biological processes including complex diseases. Machine learning provides a frequent approach to build such models. However, the models built solely from measured data often suffer from overfitting, as the sample size is typically much smaller than the number of measured features. In this paper, we propose a random forest-based classifier that reduces this overfitting with the aid of prior knowledge in the form of a feature interaction network. We illustrate the proposed method in the task of disease classification based on measured mRNA and miRNA profiles complemented by the interaction network composed of the miRNA–mRNA target relations and mRNA–mRNA interactions corresponding to the interactions between their encoded proteins. We demonstrate that the proposed network-constrained forest employs prior knowledge to increase learning bias and consequently to improve classification accuracy, stability and comprehensibility of the resulting model. The experiments are carried out in the domain of myelodysplastic syndrome that we are concerned about in the long term. We validate our approach in the public domain of ovarian carcinoma, with the same data form. We believe that the idea of a network-constrained forest can straightforwardly be generalized towards arbitrary omics data with an available and non-trivial feature interaction network. The proposed method is publicly available in terms of miXGENE system (http://mixgene.felk.cvut.cz), the workflow that implements the myelodysplastic syndrome experiments is presented as a dedicated case study.



## 1. Introduction

Onset and progression of heterogeneous multifactorial diseases depend on a combination of defected or altered genes, which is often too overly complex to be deciphered from an individual's genome only; instead it can be better manifested during the expression of genes [1]. *Gene expression* (GE) is the overall process by which information from a genome is transferred towards anatomical and physiological characteristics generally called *phenotype*. During the process, a gene is transcribed into the molecule of *messenger* RNA (mRNA), subjected to several transcription and translational regulatory mechanisms, and usually translated into a protein. The final protein level strongly afflicts the phenotype. Any dysfunction during the whole process may easily cause a disease.

The expression of a gene can be quantified as an abundance of gene transcript during its expression process. Current progress in high-throughput technologies such as microarrays and RNA sequencing enables affordable measurement of wide-scale gene expression on the *transcriptome* level. Therefore, the expression of thousands of genes can all be measured at once in each sample. One may thus feel capable of predicting disease outcome, progress or treatment response based on acquired GE data [2]. The phenotype prediction stems from the simplified assumption that a higher amount of detected mRNA implies a higher amount of translated protein, and therefore a higher manifestation of the respective gene. Phenotype prediction based on GE data is a natural learning task. However, many instances of this task become non-trivial within currently available GE data. The data are noisy and a small sample size together with an immense number of redundant features often leads to overfitting.

Gene expression can be seen as a complex dynamic process with many stages, components and regulatory mechanisms. A phenotype is not afflicted by particular genes separately, but there is a concert of genes involved in the expression process. The expression activities of genes are often indirectly linked together by interactions between respective proteins. The *protein–protein interactions* [3] may be involved in transporting and metabolic pathways, or in

* Corresponding author.
*E-mail addresses:* andelmi2@fel.cvut.cz (M. Anděl), klema@fel.cvut.cz (J. Kléma), zdenek.krejcik@uhkt.cz (Z. Krejčík).

http://dx.doi.org/10.1016/j.ymeth.2015.04.006
1046-2023/

constitution of *protein complexes*. Another component of the *gene network* are the interactions between microRNAs and their target genes [4].

MicroRNAs (miRNAs) [5] serve as a component of the complex machinery which eukaryotic organisms use to tune protein synthesis. They are short (∼21 nucleotides) noncoding RNA sequences which mediate post-transcriptional repression of mRNA via RNA-induced silencing complex (RISC), where miRNA serve as a template for recognizing complementary mRNA. The complementarity level of miRNA–mRNA binding initiates one of two possible mechanisms: the complete homology triggers *degradation* of target mRNA, whereas a partial complementarity leads to translational *inhibition* of target mRNA [6]. The level of miRNA expression can be measured by (e.g.) miRNA microarrays, analogically to mRNA profiling. The interactions between miRNAs and their target mRNAs, as well as interactions between proteins, are experimentally assessed in vitro or algorithmically predicted based on the structural properties of interacting molecules.

Since the journey from a genome to its phenotype manifestation is so complex and nontrivial, current trends in gene expression data analysis aim toward the integration of multiple measurement types from multiple stages of the gene expression process [7], acquired from the same set of tissues. Such an integrative analysis should provide a broader view of gene expression as a whole. This work extends our previous approaches to integrate traditional mRNA and miRNA measurements in the domain of myelodysplastic syndrome data based on non-negative matrix factorization with prior knowledge [8] and subtractive aggregation for deterministic models of the inhibition effect of miRNA [9]. In this paper, we propose a new method, based on random forest framework, which integrates heterogeneous omics features through the knowledge of their mutual interactions. Interlinking the features by their possible interactions improves the robustness and interpretability of resulting models, and improves their empirical validity in terms of classification accuracy.

The paper is organized as follows. Section 2 reviews the recent efforts on regularization with prior knowledge in ill-posed problems with special emphasis on omics data. Section 3 firstly describes the data domain and subsequent classification tasks. Then the method itself, designed for these classification tasks is sketched, while a way of interpreting resulting models is proposed. Next, the ovarian carcinoma domain used for validation as well as the format of employed domain knowledge is described. The methodology developed and used is deeply theoretically analyzed in Section 4. Section 5 provides experimental results in terms of empirical validity and interpretability respectively, i.e., the predictive accuracy and examples of discovered interactions along with their biological meaning. The results are then discussed in Section 6. Section 7 concludes the paper.

## 2. Related work

Learning from GE data is a challenging task due to its complexity and heterogeneity. On top of that, the number of variables $p$ greatly exceeds the number of observations $n$, we are referring to the so-called $n \ll p$ problem that leads to overfitting [10]. However, certain learning algorithms may provide promising results even in ill posed problems like this. For example, *support vector machine* (SVM) [11] is capable of dealing with a large dimensionality with sufficient generalization. However, in GE data analysis, the model itself is often just as appreciated as its output. Henceforth, SVM is more or less a black-box model, which does not provide sufficient insight. Conversely, a decision tree is easily comprehensible, but its prediction results are often weak [12]. Since GE data have a large dimensionality with few samples, there is a great number of hypotheses, often based merely on random perturbations, which can perfectly split the data into classes, but lack generalization. Counter-intuitively, even decision stumps (one-level decision trees) are overfitted as a consequence.

The way to address overfitting in general is *regularization* [13]. Regularization restrains the space of all hypotheses to improve generalization. In terms of machine learning, the trade off between bias and variance is tuned to deliberate a smaller structural risk. Besides initial dimensionality reduction, it may be implemented geometrically as in the case of margin classifiers [14], through certain hypothesis assumptions, complexity penalization or domain knowledge. We will focus on the last approach here, in which we promote such hypotheses that are in accord with the existing knowledge.

The prior knowledge-based regularization approaches are popular in the molecular biology domain; in particular, in omics data analysis. In the most general way, the domain knowledge is encoded as conditional probability in statistical relational learning [15,16], or as first-order predicates in inductive logic programming [17,18]. The advantage of these approaches is the ability to tackle the knowledge from an arbitrary domain; i.e., not only omics. However, these approaches are computationally expensive in domains with a large dimension. In omics problems where the dimension commonly exceeds $10^4$, it often implies substantial problem reduction in terms of pre-processing. An alternative way is to develop a specialized learning method dedicated to a certain domain, which stems from the domain functionality and its specific assumptions and integrates them into a learning framework. As an example of dedicated method see network regularized SVM and logistic regression, [19–21] respectively, where genes related by prior known interactions are expected to contribute *similarly* to the classification function. Among others, [22] gives an overview of recent methods for the incorporation of biological prior knowledge on molecular interactions and known cellular processes into the feature selection process to improve risk prediction of patients. Johannes et al. [23] exemplifies a tool for the incorporation of gene network data into support vector machines. Rapaport et al. [24] proposes both supervised and unsupervised learning based on spectral decomposition of gene expression profiles with respect to the eigenfunctions of the underlying gene network graph.

Regularization through domain knowledge is not such a frequent issue in the case of ensemble classifiers. The prior knowledge model and ensemble model are often regarded as two sides of the same coin, as both try to address the generalization problem and model enhancement. However, there is no reason not to combine both. Zhou et al. [25] uses gene ontology terms and miRNA–mRNA target relations to create an ensemble of centroid-based weak classifiers based on tree-like modules to forecast the prognosis of breast cancer patients. Su et al. [26] integrates linguistic knowledge into random forest language models originally based on n-gram counts only. The author illustrates the applicability of the ensembles in morphological language models of Arabic, prosodic language models for speech recognition and a combination of syntactic and topic information in language models. Dutkowski et al. [27] proposes a random forest-based method, where the building of trees is guided by a protein network. The authors proposed a procedure for the validation of network decision modules through the forest and demonstrated that the validated modules are robust and reveal causal mechanisms of cancer development. However, their search strategy most likely does not improve the classification accuracy of resulting models. Chen et al. [28] iteratively builds random forests through a weighted sampling of the variables taken from modules of correlated genes. They use the OOB (out-of-bag) importance estimate of each gene involved in the forest to adapt its weight and the weight of its module for