Network-Constrained Forest for Regularized Omics Data Classification

Michael Anděl, Jiří Kléma
Department of Computer Science
Czech Technical University
Technická 2, Prague, Czech Republic
Email: {andelmi2,klema}@fel.cvut.cz

Zdeněk Krejčík
Department of Molecular Genetics
Institute of Hematology and Blood Transfusion
U Nemocnice, Prague, Czech Republic
Email: zdenek.krejcik@uhkt.cz

***Abstract*—Contemporary molecular biology deals with a wide and heterogeneous set 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 minimizes 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.**

I. 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; it is instead 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 eventually translated into a protein. The final protein level strongly afflicts the phenotype. Any dysfunction during this process may easily cause a disease.

The expression of a gene can be quantified in terms of the measured amount of the gene transcripts during the 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* may either be involved in transporting and metabolic pathways, or in constitution of *protein complexes*. Another component of the *gene network* are the interactions between microRNAs and their target genes.

microRNAs (miRNAs) [3] serve as a component of the complex machinery which eukaryotic organisms use to tune protein synthesis of expressed genes. They are short ($\sim$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 [4]. The level of miRNA expression can be measured by (e.g.) miRNA microarrays, analogically to mRNA profiling. The resulting dataset, called the miRNA expression profile, contains, similarly to mRNA profiles, biological samples as data instances; only this time the attributes are individual miRNA sequences. 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 [5], 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 [6] and subtractive aggregation for deterministic models of the inhibition effect of miRNA [7]. 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 II reviews the recent efforts on regularization with prior knowledge in ill-posed problems with special emphasis on omics data. Section III describes the proposed concept of network-constrained forests. Section IV describes the myelodysplastic syndrome domain where the proposed methodology was applied, as well as the ovarian carcinoma domain used for validation; it also defines the learning tasks and summarizes the experimental protocol. Section V provides and discusses experimental results. Section VI concludes the paper.

II. Related Work

As mentioned above, learning from GE data is a challenging task due to its complexity and heterogeneity. On top of that, the number of variables p often exceeds the number of observations n, we are referring to the so-called $n \ll p$ problem that leads to overfitting. However, certain learning algorithms may provide promising results. For example, *support vector machine* (SVM) [8] is capable of dealing with 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 [9]. 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. Paradoxically, even decision stumps are overfitted as a consequence.

Random forest (RF) [10] addresses the generalization issue by randomly sampling a limited number of features to build an entire forest of decision trees. The resulting classifier is an *ensemble model* composed of trees as weak classifiers. The decisions of particular trees are merged to form the final decision. The diversity of the trees should enhance the generalization and stability of the model. Nevertheless, due to immense dimensionality, even in the case of a limited number of sampled features (often the square root of the original dimensionality), there is still a plethora of ways to split the data perfectly. The resulting forest of stumps is not of great diversity. Therefore, RF as a learning method often fails to substantially improve decision trees for GE data.

One way to address overfitting in general is *regularization*. Regularization restrains the space of all hypotheses to improve generalization and consequently even the overall accuracy. 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 [11], through 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 omics data analysis, in particular. Among others, [12] 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. [13] exemplifies a tool for the incorporation of gene network data into support vector machines. [14] 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. [15] 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. [16] 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. [17] proposes an RF-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. [18] 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 the sampling session in the next iteration. Using bootstrapping with subsequent OOB assumes a sufficiently large data sample.

III. Methods

In this section, we propose network-constrained forests for regularized omics data classification. Firstly, we sketch out an intuitive motivation for our proposed method. Then the method itself is formalized. Subsequently, the way in which the ensemble is analyzed to understand the interactions among features is presented.

A. Motivation

Our method is based on the general assumption that the algorithm is provided with a data matrix in which each instance is described by the values of a finite set of features. This matrix is complemented by a feature interaction network that expresses dependency among features. In the context of molecular biology, we focus on the classification of omics data where

instances correspond to biological samples and features can be transcript amount, protein content, metabolite characterization or gene methylations. The models are regularized with the aid of biological networks. In the given setting, we work mainly with transcriptomic data and employ gene regulatory networks. We claim that utilizing gene regulatory networks is advisable regarding the stability and comprehensibility of the resulting model, namely when treating the usual $n \ll p$ setting. From a biological point of view, disease is caused by a dysfunction of one or more, often multiple, genes. The defect of these *causal* genes need not be clearly detectable in their mRNA expression, but may be observable in the expression of interacting genes. Similarly, the effect of translational repression by miRNA may be observable on target mRNA expression in the case of mRNA degradation, when the mRNA molecule is wiped out entirely. While in the case of regulation by inhibition, when mRNA is only (partially) prevented from translation, the effect may be noticeable in the respective miRNA expression or is detectable in multivariate expression of involved genes and miRNAs instead.

Our intention is to sample the variables for building particular trees in the forest from sets of related biological entities rather than from the random gene sets as in the case of RF. The trees based on related entities should be: 1) diverse, since the related genes are more likely to be correlated and vice versa, 2) more accurate, as a learner built on more phenotype- or mechanism-related genes is not as liable to overfit in comparison with its counterpart built from unrelated genes.

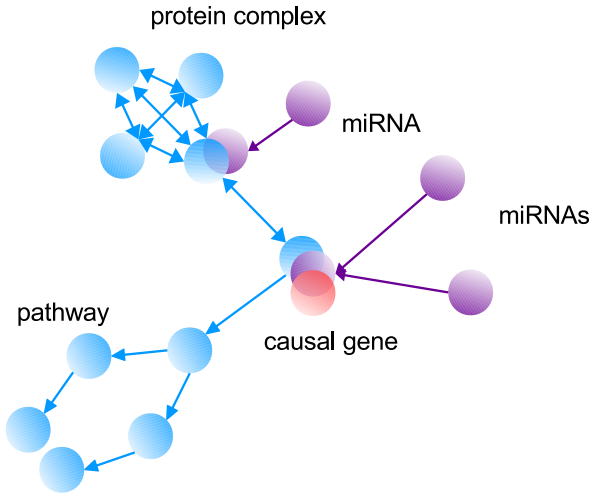


Fig. 1: Part of a gene network where a decision tree is built. Genome units are in red, transcripts and their interactions in magenta, and protein units and interactions in blue.

B. Network Constrained Forest

Our proposed method stems from the intuitive biological background mentioned above. *Network constrained forest* (NCF) learns decision trees on those features that lie close to the previously reported causal genes in the feature interaction network. This selection is unlike RF, which uses randomly selected predictors. For each node of each decision tree the RF algorithm samples the feature subset from a uniform distribution of genes. Our idea was to modify this approach by biasing the sampling process towards the genes and features in general, which have been previously reported as candidates for causing the phenomenon under study (typically a disease, treatment outcome, etc.), or towards the genes which directly or indirectly interact with those candidate genes.

1: **procedure** BUILDNCF($\mathcal{T}, \mathcal{N}, \mathcal{C}, s$)
2: $F \leftarrow array[s]$ ▷ Initialize empty forest.
3: $S \leftarrow array[s]$ ▷ Each tree has a seed.
4: $S \leftarrow sampleWithRep(\mathcal{N}, \mathcal{C})$ ▷ Get seeds.
5: **for** $t \leftarrow 1 \ldots s$ **do** ▷ Iterate over s trees.
6: $\mathcal{F} \leftarrow prWalk(\mathcal{N}, S[t])$ ▷ Get constrained set.
7: $F[t] \leftarrow buildTree(\mathcal{T}, \mathcal{F})$ ▷ Learn t-th tree.
8: **end for**
9: **return** F
10: **end procedure**

Fig. 2: Pseudocode of NCF algorithm.

A straightforward way to implement this concept is to modify the distribution that the genes are sampled from. The situation we want to deliver is as follows. The genes that are more related to a candidate causal gene in terms of gene network are more likely to be chosen. An example of such a causal gene neighborhood is displayed in Figure 1.

Let $\mathcal{G} = \{g_1, ..., g_M\}$ be the genes for which transcription level is measured, $\mathcal{R} = \{r_1, ..., r_N\}$ be miRNAs with available transcription levels. Both the measurements span the identical sample set $\mathcal{S} = \{s_1, ..., s_S\}$, the transcription matrix is $\mathcal{T} : (\mathcal{G}|\mathcal{R}) \times \mathcal{S} \rightarrow \mathbb{R}$. Let $\mathcal{P} \subset \mathcal{G} \times \mathcal{G}$ represent interactions between particular genes through respective proteins. Let $\mathcal{I} \subset \mathcal{R} \times \mathcal{G}$ represent interactions between particular miRNAs and their target genes. By merging these units and interactions we define a gene network $\mathcal{N} = (V, E)$ as a directed graph with $V = \mathcal{G} \cup \mathcal{R}$ and $E = \{(u, v) : u, v \in \mathcal{G}, (u, v) \in \mathcal{P} \vee (v, u) \in \mathcal{P}\} \cup \{(u, v) : u \in \mathcal{G}, v \in \mathcal{R}, (u, v) \in \mathcal{I}\}$. In the other words, protein-protein interactions are considered as symmetric, while miRNA-target interactions not. Finally, let $\mathcal{C} \subset \mathcal{G}$ be the genes that are believed to be associated with a disease in terms of genotype (i.e., polymorphisms, mutations).

The entire algorithm is formalized in Figure 2. As the procedure *buildTree* is taken from the RF implementation, the core of NCF obviously lies in the learning of $\mathcal{F}$ that represents a constrained set of features for construction of the t-th tree. The corresponding *prWalk* procedure that pseudorandomly samples $\sqrt{V}$ features (the common RF heuristic) from the network neighborhood of the seed feature is described in the following subsection. The procedure *sampleWithRep* implements a weighted sampling with replacement, the sampling weight is determined by the size of the neighborhood of each causal gene. The genes with a larger neighborhood appear more frequently in the resulting set of seeds (supposing that $s \gg |\mathcal{C}|$). Note that NCF does not employ bootstrapping of the sample set as it assumes a small sample size and sufficient tree decorrelation reached through constraints.

C. Building the constrained feature set

As already mentioned above, each tree is constructed from a limited set of candidate features lying around the causal seed gene. *prWalk* takes the network, the seed feature and the feature set size and proceeds as follows to deliver the neighborhood of the given size. Until the given number of unique features is reached, it starts a new random walk of length w from the seed and walks the network. All the features on the path are included in the output set. The procedure follows the traditional random walk on directed graphs. The only exception is the change of miRNA-mRNA target edge direction. First, miRNAs have to be accessible from the causal gene. Second, the edges cannot be treated as undirected since miRNAs often have a large number of targets and constitute network hubs, i.e., the high-degree nodes leading to the small-world phenomenon. To minimize the ability of reaching distant network nodes, miRNAs represent terminal nodes in the walk.

One of the key issues is finding the optimal walk length w. Short walks tend to generate small neighborhoods, the individual trees grow larger to fit the training data. Long walks get closer to the traditional RFs as the constrained feature sets become less dependent on the seed and the network topology. Note that NCF does not fully converge to RF for any walk length as the whole path starting in the seed is always taken.

In this paper we propose a heuristic to set w. The heuristic is based on a decreasing *incidence of underfitted trees* in the forest with growing walk length. In our case, the walk is stopped one step before the empirical incidence of underfitted trees I (the percentage of trees that do not fit the training data perfectly) ceases to decrease or approaches zero:

$$w = \underset{l=1\ldots10}{\arg\min}(I(l+1) < \epsilon \vee I(l) - I(l+1) < \epsilon) \quad (1)$$

At this length, the neighborhood is large enough to provide sufficient accuracy, yet still small enough not to overfit the training data. The tree depth is limited to 2 here to avoid perfect fit regardless the walk length. The value of ϵ was set to 1%. The relationship between the walk length and the underfitted tree incidence is shown in Section V and Figure 3. The value of w adjusts the learning bias to the given problem. Usually, NCF uses fewer features than RF while reaching equal or better predictive results.

D. Understanding the model

Random forests are not by far black-box models used solely for prediction. The ensembles are frequently used for variable importance estimation, feature selection or sample proximity evaluation ([19], [20], [21]). However, in the context of NCFs, variable interactions apparently represent the most interesting piece of knowledge that can be extracted from the model. In the traditional scenario that does not involve prior knowledge, the interactions are extracted purely from the measurements [22], [23]. A pair of variables is considered to interact if a split on one variable in a tree makes a split on the other variable either systematically more possible or less possible. The interactions often serve to improve the bias in variable selection stemming from variable interaction effects.

We deal with the prior set of interactions equivalent to a feature network when building the weak classifiers. Each tree belonging to the ensemble is believed to be local in terms of this feature network, i.e., to contain features that lie close to the seed gene. For this reason, we may extract the *empirical interactions* that correspond to interactions that make edges in the shortest path connecting a pair of tree neighboring nodes in the prior gene regulatory network. By searching the whole forest, we find a large number of empirical interactions that can be employed in statistical validation of the prior interaction network under the given biological conditions. The more counts a prior interaction gets in the empirical phase, the more active it seems to be in the given context. In other words, we employ the common statistical and data mining formula "data + prior knowledge $\rightarrow$ knowledge", an interaction is extracted either if it is obvious from the measurements themselves or it is contained in the prior interaction set and not invalidated by the measurements. The prior and posterior empirical interaction sets can also be compared.

IV. Experiments

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. Here we describe the experimental protocol, two particular domains and the resources employed for the building of the gene regulatory network used as prior knowledge.

A. Domain Description

The data, provided by our collaborative lab at the Institute of Hematology and Blood Transfusion in Prague, are related to *myelodysplastic syndrome* (MDS) [24]. Illumina miRNA (Human v2 MicroRNA Expression Profiling Kit, Illumina, San Diego, USA) and mRNA (HumanRef-8 v3 and HumanHT-12 v4 Expression BeadChips, Illumina) expression profiling were used to investigate the effect of lenalidomide treatment on miRNA and mRNA expression in bone marrow (BM) CD34+ progenitor cells and peripheral blood (PB) CD14+ monocytes. Quantile normalization was performed independently for both the expression sets, then the datasets were scaled to have the identical median of 1. The mRNA dataset has 16,666 attributes representing the GE level through the amount of corresponding mRNA measured, while the miRNA dataset has 1,146 attributes representing the expression level of particular miRNAs. The measurements were conducted on 75 samples labeled as follows. The sample was either healthy, or afflicted. If afflicted, it was further categorized according to genotype background as possible incidence of the long arm of chromosome 5 (del(5q) or non-del(5q)), and according to the stage of lenalidomide treatment, i.e. before treatment (BT), or during treatment (DT). Together with 2 types of tissue for each of these configurations it adds up to 10 categories.

From these categories we selected 7 binary classification tasks with a clear clinical or biological interest. The tasks were to differentiate: 1) healthy samples and afflicted samples with a particular genotype and treatment stage, 2) treatment stage of afflicted samples with del(5q), and 3) genotype background (incidence of del(5q)) of untreated samples.

B. Validation Domain

To validate our method we used data related to another genetically determined disease from an independent source. We downloaded 17,814 mRNA (Agilent 244K Custom Gene Expression G4502A-07) and 799 miRNA (Agilent Human miRNA Microarray Rel12.0) profiles with matching samples related to ovarian carcinoma (OC) from The Cancer Genome Atlas (TCGA) repository [25]. TCGA data encompasses hundreds of heterogeneous tumor samples with different clinical backgrounds. We choose the OC, which contains a sufficient number of matching mRNA and miRNA profiles, clinically well annotated. Hence, we defined two validation data sets with regards to sample homogeneity in terms of their clinical annotation and balanced class distributions. The first data set contains 58 tumor samples of high grade (G3) and late stage (Stage IV), the latter contains 64 lower grade (G2) tumor samples. The dichotomized overall survival of respective patients was chosen as a target attribute to be learned. The survival dichotomization threshold was set to the median value of overall survival, i.e. 25 months for high grade tumor dataset and 40 months for the lower grade. This preprocessing concludes with long-term and short-term survival classes, each with 29 samples for the high grade dataset and with 31 and 33 samples, respectively, for the latter dataset.

C. Domain Knowledge

Considering domain knowledge, in terms of gene networks, we downloaded the interactions between proteins, and genes and miRNAs, from the following publicly available databases. In vitro validated miRNA-mRNA interactions were obtained from TarBase 6.0 [26], while in silico predicted relations were downloaded from miRWalk database [27]. Protein-protein interactions were obtained from Human Protein Reference Database [28] and from [29], as to the experimentally validated and algorithmically predicted interactions, respectively. Eventually, we ended up with 9,077 genes involved in 79,288 protein-protein interactions, 463 miRNAs in 92,886 miRNA-target interactions. Regarding the validation domain, we handled 8,073 genes involved in 81,067 protein-protein interactions, 417 miRNAs in 84,332 miRNA-target interactions. The candidate causal genes, a total of 145 and 220 genes associated with MDS and OC respectively, were obtained from [30].

D. Experimental Protocol

To validate our method we have extended an implementation of RF in Scikit-learn [31], a Python machine learning library. Then we run number of robust experiments on our NCF. Random forest, standard classification and regression tree (CART), linear SVM and naïve Bayes (NB) classifier served as benchmark learners [32]. A simple decision tree was introduced to assess generalization of forest based algorithms. Each learning algorithm was validated in 5-times repeated 10-fold stratified cross-validation. Since all the learners should be randomly initialized, we ran each of the validation processes from 10 random seeds. It comes out to $3 \times 10 \times 5 \times\ 10 = 1500$ learning epochs. Forest based algorithms were run with 1,024 base trees.

Since we deal with classes of different sizes, we use the Mathews correlation coefficient (MCC) as a balanced quality measure. It returns a value of between -1 and +1; +1 represents a perfect match between annotation and prediction, 0 equals random prediction and 1 indicates absolute disagreement between annotation and prediction.

V. Results and Discussion

The results illustrate an empirical evaluation of our method within 9 classification tasks. The evaluation is performed in terms of the classification accuracy plotted in Figure 3. The graphs depict the progress of an empirical estimate of NCF classification accuracy as a function of walk length. Since the walk length w is a key learning parameter of NCF, we also plot the incidence of underfitted trees, which serves as a heuristic to assess a proper value of w. The MDS and ovarian cancer class definitions reached through dichotomization are available in the subfigure legends of Figure 3. The classes in the individual tasks are encoded according to the nomenclature in Section IV-A and Section IV-B, respectively.

The accuracy reached with the proper heuristic setting of w (note that it is not the walk length with the optimistically biased maximum accuracy) is compared with the other learning algorithms in Table I. The overall picture can be seen in terms of an average over 9 tasks. As the MCC values reached in different tasks can be seen as incomparable, the classification algorithms are also evaluated in terms of their average ranking. The algorithms are sorted and ranked according to their MCC scores in each of the tasks (from 1st to 4th) first, then the average of their ranks is calculated. The lower the rank, the better the algorithm. Both in Figure 3 and Table I only the median of 10 randomly seeded runs representing a robust accuracy estimate for each classification task are shown.

TABLE I: Median MCC values for 7 MDS classification tasks and 2 OC tasks. The w values set according to Equation 1 were applied for NCF. The other classifiers worked with the default settings.

task #	class ratio	NCF	RF	CART	SVM	NB
MDS1	5:10	0.54	0.58	0.26	0.56	**0.74**
MDS2	4:10	0.82	0.50	0.51	**1**	0.50
MDS3	6:10	**1**	0.82	0.57	**1**	0.82
MDS4	4:9	0.89	0.58	0.64	**1**	0.38
MDS5	6:11	0.84	0.54	0.79	**1**	0.56
MDS6	13:9	0.64	0.66	0.1	**0.66**	0.20
MDS7	5:11	**0.78**	0.27	0.13	0.34	0.20
OC1	29:29	**0.37**	0.32	0.1	0.32	0.23
OC2	31:33	**0.49**	0.383	0.25	0.36	0.46
Average accuracy		**0.71**	0.52	0.37	0.69	0.42
Average ranking		**2**	3.1	4.1	**2**	3.2

The table illustrates a shift in predictive performance among several learning methods, starting from the comprehensible but empirically invalid decision tree through random forest to the accurate though black-box SVM. The results suggest that our network-enriched RF provides a good trade off between these two extremes. NCF shows good classification accuracy, while being more comprehensible than black-box models (see Section V-A). In most of the cases, NCF has a better or equal predictive power than the state-of-the-art RF and as a whole, in terms of classification accuracy, is even competitive with the black-box SVM.

In order to gain a deeper insight into the mechanism of NCF, in Figure 3 the relationship between the empirical estimate of generalization error, walk length as a regularization parameter, and incidence of underfitted trees used to set a heuristic value of the parameter can be observed. This relationship shows a few consistent patterns. In MDS1 (Figure 3a) a prematurely converging heuristic can be observed. This most probably leads to underfitting of the model, which leads to the stagnation of predictive accuracy close to the baseline represented as a performance of RF. The mutually related tasks MDS2 and MDS4 (Figure 3b and 3d), which share the definition of one the classification classes, show the heuristic falling to zero; at the same time, the initially promising value of MCC is reduced. Such a trend suggests overfitting, most probably due to the small sample size of the shared class, which contains only 4 samples (see Table I). Conversely, stable good classification performance is manifested within another two mutually related tasks MDS3 and MDS5 (Figure 3c, Figure 3e). Very interesting results are shown in task MDS7, which manifest a slow decrease of the heuristic and attendant growth of the accuracy far above the baseline. As to the OC tasks which are obviously more complex due to the larger number of samples (see Table I), NCF, under the proper parametrization, beats standard RF and even the black-box SVM. As to the heuristic settings of the walk length parameter w, it can be stated that the heuristic finds the values of generalization error very close to their optima.

A. NCF interaction exploitation

To illustrate the process of NCF understanding, the set of 10 differently initialized forests constructed for MDS7 was taken. The task is to assess the impact of treatment in the group of BM del(5q) patients. 107 feature interaction pairs that appeared 10 and more times were extracted.

For several interaction pairs, their hypothetically calculated relationship or involvement in MDS and/or leukemia have a solid experimental support in reality. One example with a high score is an interaction pair found in the case of EGFR–CBL. Whereas CBL, the E3 ubiquitin-protein ligase involved in cell signalling and protein ubiquitination, is already known to control the fate of EGRF (epidermal growth factor receptor) ([33]), mutations in *Cbl gene* have been related to MDS and acute myeloid leukemia (AML) ([34], [35]). Additionally, CBL forms another interesting pair, e.g., the interaction between CBL and ABL1 (Abelson murine leukemia viral oncogene homolog 1) is also well documented (e.g., [36]). *Abl1* is a proto-oncogene that, activated by t(9;22) translocation, creates a new fusion gene, BCR-ABL which is typically associated not only with chronic myelogenous leukemia (CML) but also in some cases with acute lymphoblastic leukemia (ALL) and occasionally with AML ([37]). Still more pairs contain CBL; from those, at least interactions with CRK, CD2AP and AXL were previously reported ([38], [39], respectively), although up to now they have not been seen as relevant factors in MDS or leukemia.

Another important high-score hit is NPM1–RAD50. Besides functioning as DNA-repair proteins found to be deregulated together in ovarian cancer ([40]), NPM1 (nucleophosmin) is known to be involved in AML, MDS, and acute promyelocytic leukemia ([41]). Interestingly, RAD50, as a DNA-repair protein, is also a potential factor in the etiology of AML and possibly MDS ([42]).

Besides the aforementioned genes, interaction pairs involve other genes, e.g., KRAS, JAK3, STAT3, and SYK, whose occurence in MDS is known and under further investigation (for an overview, see [30]). Somewhat surprisingly, within the interactions there are only several hits for miRNA-coding genes previously reported as deregulated in MDS (e.g. miR-124, miR-329, miR-355 and miR-155) (reviewed in [43]). The other miRNA listed in interactions could then possibly represent new targets to which we should turn our attention considering MDS, i.e. miR-495 which has previously been associated with acute myeloid leukemia with mixed lineage leukemia rearrangements [44] but not with MDS.

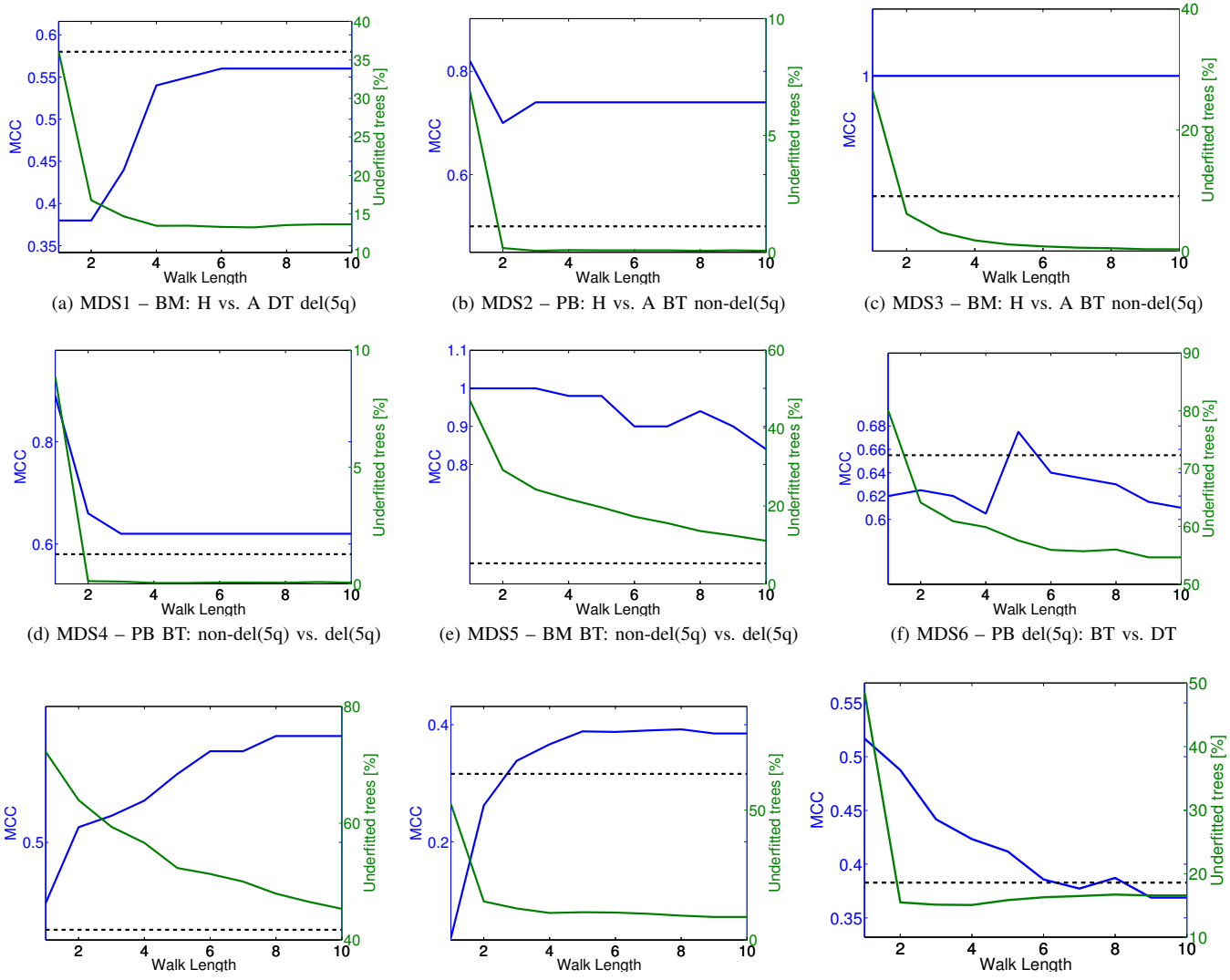
To sum up, 36 out of 107 (33 %) high-scoring interaction pairs contain a gene (or the whole pair) whose (more or less significant) relationship to MDS has already been reported [30]. Such an occurrence is far from random and is obviously the result of a successful computational approach. Therefore, some of the interactions (or interacting counterparts) could turn out to be promising candidates for further investigation for their role in MDS and/or leukemia through both a literature search as well as laboratory experiments.

VI. Conclusion and Future Work

We proposed a general parameter-free method for learning from high-dimensional and low-sample size data complemented by a feature interaction network. The method of network-constrained forest stems from the well-known random forests. The main difference is that decorrelation of the individual weak classifiers is not reached through bootstrap sampling and random subsetting of the features, but pseudorandom subsetting driven by the feature interaction network. The individual trees deal with feature sets sampled from different areas of a feature network, the curated feature interactions thus tend to be promoted. Still, they are not strictly imposed; the method remains stochastic in its nature and an arbitrary feature relationship may appear in a tree. The probability of its occurrence increases with the decreasing path length between the pair of features in the network and increasing interaction observed in measurements. Unlike our previous efforts, we do not rely on feature extraction based on prior modules such as pathways or simple interaction subgraphs [45], [6], [7].

The method was applied to improve classification accuracy and comprehensibility of gene expression-based disease models. The obtained results suggest that introducing domain knowledge improves the accuracy of the forest and increases its compliance with the current knowledge. We believe that the method is able to benefit from stochastic identification of the subset of earlier reported general interactions; this subset manifests in the given context.

In future work, we will aim at truly omics experiments. We will employ more data types such as epigenetic data, namely DNA methylation arrays, and further extend utilized prior knowledge; for example, with information about transcription factor interactions and problem related pathways. At the moment, the main limitation lies in the simplifying assumption of measurement completeness; all the measurements must be available for all the samples, which is often not the case.



(a) MDS1 – BM: H vs. A DT del(5q)
(b) MDS2 – PB: H vs. A BT non-del(5q)
(c) MDS3 – BM: H vs. A BT non-del(5q)
(d) MDS4 – PB BT: non-del(5q) vs. del(5q)
(e) MDS5 – BM BT: non-del(5q) vs. del(5q)
(f) MDS6 – PB del(5q): BT vs. DT
(g) MDS7 – BM del(5q): BT vs. DT
(h) OC1 – G3 Stage IV: long- vs. short-term
(i) OC2 – G2: long- vs. short-term

Fig. 3: Experimental results for 7 MDS classification tasks (graph 3a – 3g) and 2 validation tasks related to the overall survival of ovarian cancer (graph 3h – 3i). The development of Mathews correlation coefficient (MCC) (in blue) and the incidence of underfitted trees (in green) with increasing walk length. The MCC values of benchmarking RFs (that do not work with the walk length) are shown in dotted lines.

Some patient mRNA profiles may missing, even though protein levels might be available for them, etc. The authors of [46], [47] demonstrate that feature networks represent a suitable regularization tool in other domains, as well; such as document topic prediction and click prediction. Eventually, we plan to proceed further beyond classification, namely to analyze the resulting forest. To be more precise, an analysis of successful trees in terms of gene ontology terms could be provided. Dealing with artificially generated data should answer the general applicability of the given method. The analysis should also work with different ratios of n and p (as the number of samples grows the methods such as sparse SVM seem to be natural competitors [48], at least in terms of accuracy) and feature network sizes and topologies as well as feature interaction strengths (the stronger the curated interactions manifest in the measurements, the more prior knowledge applies to the given domain but it is easier to be identified from the measurements themselves).

Acknowledgment

This work was supported by the grants NT14539, NT14377 and NT13847 of the Ministry of Health of the Czech Republic.

References

[1] C. Giallourakis, C. Henson, M. Reich *et al.*, "Disease gene discovery through integrative genomics," *Annu Rev Genom Hum G*, vol. 6, no. 1, pp. 381–406, 2005.

[2] T. R. Golub, D. K. Slonim, P. Tamayo *et al.*, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[3] R. C. Lee, R. L. Feinbaum, and V. Ambros, "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14." *Cell*, vol. 75, no. 5, pp. 843–54, 1993.

[4] M. R. Fabian and N. Sonenberg, "The mechanics of miRNA-mediated gene silencing: a look under the hood of miRISC," *Nat Struct Mol Biol*, vol. 19, no. 6, pp. 586–93, Jun. 2012.

[5] S. Sass, F. Buettner, N. Mueller, and F. Theis, "A modular framework for gene set analysis integrating multilevel omics data," *Nucleic Acids Res*, vol. 41, no. 21, pp. 9622–9633, Nov. 2013.

[6] M. Anděl, J. Kléma, and Z. Krejčík, "Integrating mRNA and miRNA expressions with interaction knowledge to predict myelodysplastic syndrome," in *ITAT 2013: Workshop on Bioinformatics in Genomics and Proteomics*, 2013, pp. 48–55.

[7] J. Kléma, J. Zahálka, M. Anděl, and Z. Krejčík, "Knowledge-based subtractive integration of mRNA and miRNA expression profiles to differentiate myelodysplastic syndrome," in *Proceedings of Int. Conf. on Bioinformatics Models, Methods and Algorithms*, 2014, pp. 31–39.

[8] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach Learn*, vol. 46, pp. 389–422, March 2002.

[9] M. H. Asyali, D. Colak, O. Demirkaya, and M. S. Inan, "Gene expression profile classification: A review," *Current Bioinformatics*, vol. 1, pp. 55–73, 2006.

[10] L. Breiman, "Random forests," *Mach Learn*, pp. 5–32, 2001.

[11] A. J. Smola and P. J. Bartlett, Eds., *Advances in Large Margin Classifiers*. Cambridge, MA, USA: MIT Press, 2000.

[12] C. Porzelius, M. Johannes, H. Binder, and T. Beißbarth, "Supporting information leveraging external knowledge on molecular interactions in classification methods for risk prediction of patients." *Biometrical Journal*, vol. 53, no. 2, pp. 190–201, 2011.

[13] M. Johannes, H. Frhlich, H. Sltmann, and T. Beibarth, "pathClass: an R-package for integration of pathway knowledge into support vector machines for biomarker discovery." *Bioinformatics*, vol. 27, no. 10, pp. 1442–1443, 2011.

[14] F. Rapaport, A. Zinovyev, M. Dutreix *et al.*, "Classification of microarray data using gene networks." *BMC Bioinformatics*, vol. 8, 2007.

[15] X. Zhou, J. Liu, X. Ye *et al.*, "Ensemble classifier based on context specific mirna regulation modules: a new method for cancer outcome prediction." *BMC Bioinformatics*, vol. 14, no. S-12, p. S6, 2013.

[16] Y. Su, "Knowledge integration into language models: A random forest approach," Ph.D. dissertation, The Johns Hopkins University, Baltimore, Maryland, 4 2009, 90 p.

[17] J. Dutkowski and T. Ideker, "Protein networks as logic functions in development and cancer." *PLoS Comput Biol*, vol. 7, no. 9, 2011.

[18] Z. Chen and W. Zhang, "Integrative analysis using module-guided random forests reveals correlated genetic factors related to mouse weight," *PLoS Comput Biol*, vol. 9, no. 3, p. e1002956, 03 2013.

[19] A. Verikas, A. Gelzinis, and M. Bacauskiene, "Mining data with random forests: A survey and results of new tests," *Patt Recogn*, vol. 44, no. 2, pp. 330 – 349, 2011.

[20] H. Deng and G. Runger, "Gene selection with guided regularized random forest," *Patt Recogn*, vol. 46, no. 12, pp. 3483 – 3489, 2013.

[21] M. B. Kursa, "Robustness of random forest-based gene selection methods," *BMC Bioinformatics*, vol. 15, no. 1, p. 8, 2014.

[22] K. Lunetta, L. B. Hayward, J. Segal, and P. Van Eerdewegh, "Screening large-scale association study data: exploiting interactions using random forests," *BMC Genetics*, vol. 5, no. 1, p. 32, 2004.

[23] C. Strobl, A.-L. Boulesteix, T. Kneib *et al.*, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, no. 1, p. 307, 2008.

[24] A. Vašíková, M. Běličková, E. Budinská, and J. Čermák, "A distinct expression of various gene subsets in cd34+ cells from patients with early and advanced myelodysplastic syndrome." *Leuk Res*, vol. 34, no. 12, pp. 1566–72, 2010.

[25] (2014) The TCGA Research Network. [Online]. Available: http://cancergenome.nih.gov/

[26] T. Vergoulis, I. S. Vlachos, P. Alexiou *et al.*, "TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support." *Nucleic Acids Res*, vol. 40, pp. D222–9, 2012.

[27] H. Dweep, C. Sticht, P. Pandey, and N. Gretz, "miRWalk - Database: Prediction of possible miRNA binding sites by "walking" the genes of three genomes." *J Biomed Inform*, vol. 44, no. 5, pp. 839–47, 2011.

[28] T. S. Prasad, R. Goel, K. Kandasamy *et al.*, "Human protein reference database - 2009 update." *Nucleic Acids Res*, vol. 37, no. Database-Issue, pp. 767–772, 2009.

[29] A. Bossi and B. Lehner, "Tissue specificity and the human protein interaction network." *Mol Syst Biol*, vol. 5, no. 1, Apr. 2009.

[30] W. Yu, M. Clyne, M. J. Khoury, and M. Gwinn, "Phenopedia and genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations." *Bioinformatics*, vol. 26, no. 1, pp. 145–146, 2010.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort *et al.*, "Scikit-learn: Machine learning in Python," *J Mach Learn Res*, vol. 12, pp. 2825–2830, 2011.

[32] T. Hastie, R. Tibshirani, J. Friedman *et al.*, *The elements of statistical learning*. Springer, 2009, vol. 2, no. 1.

[33] G. D. Visser Smit, T. L. Place, S. L. Cole *et al.*, "Cbl controls egfr fate by regulating early endosome fusion," *Science signaling*, vol. 2, no. 102, p. ra86, 2009.

[34] J. Rocquain, N. Carbuccia, V. Trouplin *et al.*, "Combined mutations of asxl1, cbl, flt3, idh1, idh2, jak2, kras, npm1, nras, runx1, tet2 and wt1 genes in myelodysplastic syndromes and acute myeloid leukemias." *BMC Cancer*, vol. 10, p. 401, 2010.

[35] M. Naramura, S. Nadeau, G. A. Bhopal Mohapatra *et al.*, "Mutant cbl proteins as oncogenic drivers in myeloproliferative disorders," *Oncotarget*, vol. 2, no. 3, p. 245, 2011.

[36] T. Miyoshi-Akiyama, L. M. Aleman, J. M. Smith *et al.*, "Regulation of cbl phosphorylation by the abl tyrosine kinase and the nck sh2/sh3 adaptor," *Oncogene*, vol. 20, no. 30, pp. 4058–4069, 2001.

[37] M. Talpaz, N. P. Shah, H. Kantarjian *et al.*, "Dasatinib in imatinib-resistant philadelphia chromosome–positive leukemias," *New England Journal of Medicine*, vol. 354, no. 24, pp. 2531–2541, 2006.

[38] M. Garcia-Guzman, E. Larsen, and K. Vuori, "The proto-oncogene c-cbl is a positive regulator of met-induced map kinase activation: a role for crk adaptor." *Oncogene*, vol. 19, no. 35, pp. 4058–4065, 2000.

[39] P. Valverde, "Effects of gas6 and hydrogen peroxide in axl ubiquitination and downregulation," *Biochem Biophys Res Commun*, vol. 333, no. 1, pp. 180–185, 2005.

[40] R. S. Kalra and S. A. Bapat, "Enhanced levels of double-strand dna break repair proteins protect ovarian cancer cells against genotoxic stress-induced apoptosis," *J Ovarian Res*, vol. 6, no. 1, p. 66, 2013.

[41] B. Falini, I. Nicoletti, N. Bolli *et al.*, "Translocations and mutations involving the nucleophosmin (npm1) gene in lymphomas and leukemias," *Haematologica*, vol. 92, no. 4, pp. 519–532, 2007.

[42] J.-Y. Shi, Z.-H. Ren, B. Jiao *et al.*, "Genetic variations of dna repair genes and their prognostic significance in patients with acute myeloid leukemia," *Int J Cancer*, vol. 128, no. 1, pp. 233–238, 2011.

[43] G. Rhyasen and D. Starczynowski, "Deregulation of micrornas in myelodysplastic syndrome," *Leukemia*, vol. 26, no. 1, pp. 13–22, 2012.

[44] X. Jiang, H. Huang, Z. Li *et al.*, "mir-495 is a tumor-suppressor microrna down-regulated in mll-rearranged leukemia," *Proceedings of the National Academy of Sciences*, vol. 109, no. 47, pp. 19 397–19 402, 2012.

[45] M. Holec, J. Kléma, F. Železnỳ, and J. Tolar, "Comparative evaluation of set-level techniques in predictive classification of gene expression samples," *BMC Bioinformatics*, vol. 13, no. Suppl 10, p. S15, 2012.

[46] T. Sandler, J. Blitzer, P. P. Talukdar, and L. H. Ungar, "Regularized learning with networks of features," in *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds. Curran Associates, Inc., 2009, pp. 1401–1408.

[47] D. Chakrabarti and R. Herbrich, "Speeding up large-scale learning with a social prior," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '13. New York, NY, USA: ACM, 2013, pp. 650–658.

[48] J. Bi, K. Bennett, M. Embrechts *et al.*, "Dimensionality reduction via sparse support vector machines," *J Mach Learn Res*, vol. 3, pp. 1229–1243, Mar. 2003.