Jalali and Pfeifer *BMC Genomics* (2016) 17:501
DOI 10.1186/s12864-016-2647-9

BMC Genomics

METHODOLOGY ARTICLE

Open Access

Interpretable per case weighted ensemble method for cancer associations

Adrin Jalali[*] and Nico Pfeifer

Abstract

Background: Molecular measurements from cancer patients such as gene expression and DNA methylation can be influenced by several external factors. This makes it harder to reproduce the exact values of measurements coming from different laboratories. Furthermore, some cancer types are very heterogeneous, meaning that there might be different underlying causes for the same type of cancer among different individuals. If a model does not take potential biases in the data into account, this can lead to problems when trying to predict the stage of a certain cancer type. This is especially true when these biases differ between the training and test set.

Results: We introduce a method that can estimate this bias on a per-feature level and incorporate calculated feature confidences into a weighted combination of classifiers with disjoint feature sets. In this way, the method provides a prediction that is adjusted for the potential biases on a per-patient basis, providing a personalized prediction for each test patient. The new method achieves state-of-the-art performance on many different cancer data sets with measured DNA methylation or gene expression. Moreover, we show how to visualize the learned classifiers to display interesting associations with the target label. Applied to a leukemia data set, our method finds several ribosomal proteins associated with the risk group, which might be interesting targets for follow-up studies. This discovery supports the hypothesis that the ribosomes are a new frontier in genadaptivelearninge regulation.

Conclusion: We introduce a new method for robust prediction of phenotypes from molecular measurements such as DNA methylation or gene expression. Furthermore, the visualization capabilities enable exploratory analysis on the learnt dependencies and pave the way for a personalized prediction of phenotypes. The software is available under GPL2+ from https://github.com/adrinjalali/Network-Classifier/tree/v1.0.

Keywords: Machine learning, Cancer biomarkers, Supervised prediction, Ensemble methods, Support vector machines, Gaussian processes

Background

Over the past few decades, biology has transformed into a high throughput research field, both in terms of the number of different measurement techniques as well as the amount of variables measured by each technique (e.g., from Sanger sequencing to deep sequencing), and is more and more targeted to individual cells [1]. This has led to an unprecedented growth of biological information. Consequently, techniques that can help researchers find important insights into the data are becoming increasingly important. Predicting survival of cancer patients based on measurements from microarray experiments has been a field of great interest, but there is often very little overlap between the important genes or biomarkers identified by different studies [2]. Several reasons have been suggested to explain these findings (e.g., heterogeneity of cancer samples or insufficient sample size). Attempts have been made to incorporate additional information from other sources, such as protein-protein interaction (PPI) networks, to make the predictions more robust [3]. One of the latest approaches integrates network and expression data by introducing a network-induced classification kernel (NICK) [4]. Although this method exhibits state-of-the-art performance, the way it penalizes genes that are connected to not-predictive genes can result in selection of isolated features as important features for prediction. We observed this bias of the method towards isolated

*Correspondence: ajalali@mpi-inf.mpg.de
Department of Computational Biology and Applied Algorithmics, Max Planck Institute for Informatics, Campus E1 4, 66123 Saarbrücken, Germany

nodes on additional experiments on synthesized data as shown in Additional file 1. Another issue is that in PPI networks, genes or proteins, which have been known to researchers longer and are well-known, are studied more and therefore have more edges connected to them; whereas less well-known genes and proteins are in sparser areas of the network. This bias might further affect the judgment of methods like NICK that use a PPI networks as an input. Consequently, we rely on the fact that such networks exist between genes and proteins, but we do not take them as input. If there is a dependence between input features, which is the case in many biological settings, our method can benefit from this effect. Otherwise, it is reduced to a standard ensemble method. Furthermore, a central assumption underlying many methods is that all data are drawn from the same unknown underlying distribution. This may not be the case, especially for heterogeneous cancer samples, and in particular not for all measured genes.

In this work, we introduce a method that is aware of this potential bias and utilizes an estimate of the differences during the generation of the final prediction method. For this, we introduce a set of sparse classifiers based on $L1$-SVMs [5], where each set of features used by one classifier is disjoint from the selected feature set of any other classifier. Furthermore, for each feature chosen by one of the classifiers, we introduce a regression model that uses additional features and is based on Gaussian process regression. These regression models are then used to estimate how predictable the features of each classifier are for each test sample. This information can then be used to find a confidence weighting of the classifiers, i.e. up-weighting classifiers with high confidence and down-weighting classifiers with lower confidence, for each test sample. Schapire and Singer show that incorporating confidences of classifiers can improve the performance of an ensemble method [6]. However, in their setting, confidences of classifiers are estimated using the training data and are thus fixed for all test samples, whereas in our setting, we estimate confidences of individual classifiers per given test sample. Another related work includes mixture of experts, in which the model trains a set of neural networks and uses a gating network to set the weights of the networks [7]. One issue with their method is that neural networks with lower performance will not be optimized as much as networks with better performance on training data since the gate module down-weights the error propagated to them. Also training of the gating network is interconnected with the neural network experts and afftects training of those modules. Our method, in contrast, trains each module independently using all training samples, and their reliability does not affect how they are trained. Bayesian hierarchical mixtures of experts takes a more similar approach, but the method is complex, and it has a high time complexity to train the architecture of the hierarchy [8].

We show that this method exhibits state-of-the-art performance for different cancer types, with gene expression or methylation data sets as the input. Since the weighting of the classifiers is customized for each test sample, the estimated confidences can offer insights into the specific characteristics of each individual's cancer. To facilitate interpretation of the model, we then create a visualization of the important genes found through this analysis for each test sample. Additionally, we show how the important genes of the training set can be found using our learning method and cross validation.

Our idea might resemble ensemble feature selection, which involves aggregating multiple feature scores from several scoring mechanisms. These scoring mechanisms vary from being several different methods, to being the same method applied to different parts of the data such as a random cross validation scheme [9]. This idea has been studied further by other researchers and they introduced two different methods to aggregate scores from different models. They use an ensemble of support vector machines which on its own has been used to select features in a given data set in other works [10]. Although we use an ensemble of support vector machines, our goal is not to give a ranking to features of the data set, rather to find multiple parsimonious gene sets that are predictive of the outcome on their own, and use all of them in parallel to predict the outcome.

Similar to this approach, in another work, iRDA uses a different approach and can report multiple parsimonious gene sets [11]. One significant difference between iRDA and our work is that we have an embedded prediction approach using these sets, which iRDA lacks. Furthermore, gene sets are somehow ordered in iRDA according to their "strength", and within each set, redundant genes are removed. In our model redundant genes can be included in two different ways. One is within different individual learners. For example, if genes g_1 and g_2 are both strong but redundant, individual learner 1 might include g_1, and individual learner 2 might include g_2. Also, if there are more redundant or related genes in the gene pool, they will be used to estimate how reliable g_1 and g_2 are. Therefore instead of dismissing them, we exploit the fact that they exist.

Related to sorting genes and testing for significance of a reported gene set, Gene Set Enrichment Analysis (GSEA) and its modifications are a commonly used tool [12, 13]. GSEA based methods rank genes depending on how much they relate to the outcome. The choice of relationship is rather free and can vary from Pearson correlation to mutual information. Then for a given gene set, a p-value is calculated by estimating how often a random gene set appears before the given set on the list. There have been several modifications and improvements to the method [14, 15]. Although it is true that GSEA is used to assess the relevance or importance of a given set to the outcome, we need to remember that a particular gene set might consist of genes that are not necessarily important on their own, but are predictive once considered together. Our method does not consider genes individually whereas GSEA does to sort the genes in the first place. Therefore we believe GSEA based methods are not suitable to assess how well our method performs.

Methods

Materials

Data sources

In this article, our method is applied to two different data types: gene expression data and DNA methylation data, which we retrieved from The Cancer Genome Atlas (TCGA) [16]. TCGA is a joint effort of the National Cancer Institute and the National Human Genome Research Institute to advance the understanding of the molecular basis of cancer. They provide access to the different measurements from cancer samples that have been analyzed to external researchers. Samples are categorized according to diagnosed cancer from which we use the following groups:

- *Acute Myeloid Leukemia (LAML)* [17]: At the time of writing, the data set includes 200 samples. 194 samples contain methylation data and we use the part of the data measured by JHU-USC HumanMethylation450 arrays. 173 samples contain mRNA data measured by HG-U133 arrays. In this article the methylation data is referred to as TCGA-LAML. Among available characteristics of samples, "risk group" and "vital status" are chosen as target classes. These labels show the aggressiveness of the disease. In our analysis, regarding risk group, {favorable} and {intermediate/normal, poor} samples form our two group, and in the analysis of vital status, {alive} and {dead} samples form our two groups of samples.
- *Breast invasive carcinoma (BRCA)* [18]: This data set includes 993 samples with clinical data, and we use the methylation data component measured by JHU-USC HumanMethylation450 arrays. Only very few samples in this data set are indicated as having metastasized (8 samples). Hence the data are analyzed according to "tumor size", "affected nearby lymph nodes", "stage", and "estrogen receptor". Estrogen receptor was shown to be an important factor in prognosis [19], and along with other factors directly affects the decision for therapy [20, 21]. For tumor size {T1, T2} samples are one category and {T3, T4} the other category; in order to analyze affected nearby lymph nodes, {N0} is compared to {N1, N2, N3}; stage is analyzed as having {stage I, stage II} vs. {stage III} samples. Estrogen receptor status of samples is either positive or negative, and they form our two classes.

Data preprocessing

To prepare gene expression data for analysis, microarray probes are mapped to their respective gene. If there are multiple probes for a gene, the median reported gene expression value of those probes is adopted as the gene expression for that gene.

Preparing the methylation data, we use the nearby gene for each methylation site available for each sample and each methylation site. The median beta value of methylation sites mapped to each gene is taken as the methylation value of the corresponding gene. In this process only methylation sites located on the promoter region of a gene are considered and others are discarded.

Boosting

For a given prediction problem the idea of boosting is to find an optimal combination of classifiers, also called "weak learners" [22]. There are many methods of finding the optimal combination of such weak learners, two of which are stochastic gradient boosting [23] and AdaBoost [24]. Stochastic gradient boosting tries to estimate the gradients of the loss function and train each individual weak learner in a way that best improves the loss function. AdaBoost tries to identify samples among given data samples that are harder to classify, and gives them more weight in the process of training individual weak learners. One way of improving AdaBoost is to take into account the confidences of predictions given by weak learners if possible and use estimated confidences in the voting process [6].

Learning a mixture of disjoint classifiers

When dealing with cancer, we need to consider the fact that tumors of the same type of cancer can be very different in nature and they are usually classified as different cancer subtypes. In fact, even one single tumor can be very heterogeneous [25]. This means that the malignancies causing the cancer to happen are genetically different between subtypes, or even within subtypes, and it is possible to have multiple underlying cellular processes causing a particular cancer. Also it is important to note that the nature of our given data is such that the input features are properties measured from genes, e.g. gene expression or methylation values, and these variables are correlated and statistically dependent on each other. Our method tries to exploit these properties of the problem to infer an interpretable model with state-of-the-art performance.

Our method can be characterized by the following key parts:

Training phase:

- Fit several individual classifiers to the data, in such a way that the features of the data they use are disjoint sets.

Prediction phase:

- Calculate the prediction confidence of each individual classifier by:
  - Estimating the reliability of input features of the classifier;
  - Estimating the confidence of the output based on the decision values.
- Calculate a weighted prediction label based on the individual classifier confidences.

Properties of the individual classifiers

A wide variety of classifiers is possible within our framework. One requirement is that the classifier is regularized (i.e., the stronger the regularization, the less complex the model gets and consequently the less features are used). The classifier is also required to report the probability of its calculated output, or to give a decision value according to which it chooses the predicted class. We use an $L1$ regularized SVM for this purpose with a linear kernel [5]. The $L1$ regularization makes the SVM sparse, i.e. using only a few input features, and the linear kernel allows us to infer which features are used in the decision function of the SVM after it is fit to the data.

Training the individual classifiers

The model starts with no individual classifiers and an empty set of excluded features. In each step, the excluded set of features is removed from the data, then a classifier is fit to the data. Next the features used by the most recent trained classifier are added to the excluded set. In the case of a linear kernel SVM, this is achieved by finding features with a non-zero coefficient in the model. This way the features being used by classifiers are disjoint and might represent different underlying causes of groups into which samples are to be classified.

Combining classifiers by estimating confidences of individual predictors

Given a set of classifiers, the question is how to combine them to come up with a joint prediction value for each test sample for which we want to predict the output label. The intuition behind combining the classifiers is to put more weight on classifiers that use features whose behavior is similar to the training data. This is motivated by the fact that some parts of the test data might behave very differently to the training data, meaning that a classifier using these features should have lower performance than a classifier using features that are distributed similarly to the training data. Therefore we need to evaluate the reliability of the input features of each individual classifier. In scenarios like gene expression or methylation analysis, we usually have many input features. Furthermore, many features are correlated and statistically dependent. The idea of our new method is to build separate prediction models for each feature of each classifier. These prediction models can then be used to obtain a confidence for the feature in a given test sample. These confidences can then be combined for each classifier to give a weighting of the classifiers for the given test sample. To evaluate an observed feature *f*, we try to choose a few statistically dependent features, and fit a model to predict *f*. To find these features, first the estimated maximal information coefficient (MIC) of all other features with feature *f* is calculated [26]. Then, features having MIC value within the top 5 % or the 5 features with highest MIC with f (if the top 5 % features consist of less than 5 features), are selected as predictors of *f*. Given a test sample, the closer the predicted value of f is to the observed value, the more reliable it is. To quantify this, we need to not only know the predicted value of the feature, but also a confidence interval for that prediction. This can be achieved using Gaussian processes, which give the mean and variance of the posterior probability under the condition of observed values for selected features. A weighted average of these values gives us the overall reliability of the features of an individual classifier.

In addition to the confidence in the classifier estimated by looking at the confidences of its individual features, we also account for the confidence that the classifier has in the prediction label of the test sample. If the method supplies such a confidence value (e.g., Gaussian processes), we can directly use it. Otherwise, we estimate it using the decision value. In our setting, the linear SVM gives a decision value whose sign defines the predicted class. Using these values we estimate a confidence for each individual classifier. Several approaches exist for deriving a confidence from the decision values [27]. Whether these or other additional methods could lead to further improvements of our method, will be topic of further study.

More formally speaking, define X to be the set of input samples, X_s to be the input vector of sample s, y_s and $\hat{y}_s$ to be respectively the original label and predicted output of sample s, Δ to be the set of individual classifiers, l_i to be an individual classifier, Φ_{l_i} the set of input features of classifier l_i, $l_i(X_s)$ to be the label predicted by classifier l_i for sample X_s, and *f* to be a feature, $X_{s,f}$ to be the observed value of feature *f* in sample X_s, $|w_{l_i}(f)|$ to be the absolute value of the weight of feature *f* in the decision function of classifier l_i, and g_f to be the Gaussian process predicting feature f using feature set Φ_f. Also $\mu_{g_f(X_s)}$ and $\sigma_{g_f(X_s)}$ are the mean and standard deviation of the posterior probability given by Gaussian process g_f under the condition of observing values of features in Φ_f, and μ_{l_i} and σ_{l_i} are respectively the expected mean and standard deviation of the decision value of classifier l_i. Here F is the cumulative distribution function of a standard normal distribution.

The training phase of the model is shown in Fig. 1, in which, N is the number of individual learners to be included in the model, Φ_l is the union over all Φ_{l_i} and $X_{-\Phi_l}$ is the input X after discarding all features of the set Φ_l. TOP is the function which selects the maximum of the top 5 and top 5 % features f' of all features ordered by MIC with feature f.

Now given a test sample X_s, the estimated confidence of a feature f is:

$$c_f(X_s) := 2 \cdot F\left(-\left|\frac{X_{s,f} - \mu_{g_f(X_s)}}{\sigma_{g_f(X_s)}}\right|\right) \qquad (1)$$

Then the overall feature reliability or confidence of a classifier l_i is estimated as:

$$c^1_{l_i}(X_s) := \frac{\sum_{f \in \Phi_{l_i}} c_f(X_s) \cdot \left|w_{l_i}(f)\right|}{\sum_{f \in \Phi_{l_i}} \left|w_{l_i}(f)\right|} \qquad (2)$$

Also the estimated output confidence of the classifier l_i is:

$$c^2_{l_i}(X_s) := 1 - 2 \cdot F\left(-\left|\frac{l_i(X_s) - \mu_{l_i}}{\sigma_{l_i}}\right|\right) \qquad (3)$$

and the final confidence of the classifier l_i is then:

$$c_{l_i}(X_s) := c^1_{l_i}(X_s) \cdot c^2_{l_i}(X_s) \qquad (4)$$

Finally, the predicted class $\hat{y}_s$ is calculated as the sign of a weighted vote among individual classifiers:

$$\hat{y}_s := \text{sign}\left(\frac{\sum_{l_i \in \Delta} c_{l_i}(X_s) \cdot l_i(X_s)}{\sum_{l_i \in \Delta} c_{l_i}(X_s)}\right) \qquad (5)$$

Visualization of model predictions

The interpretation of the model can be understood on two different ways. First we assume for a given training data set, the model is trained and a new test sample is given. For the given test sample it is possible to visualize the reliability of each used feature in individual classifiers, as well as the overall confidence of each individual classifier. Used features can be superimposed onto a PPI network as well as their reliability and the confidence of their respective individual classifier.

Gene expression and methylation level measurements from cancer samples are usually very noisy. Furthermore, cancers are usually very heterogeneous. Additionally, there might be different subgroups for each interesting group (e.g., cancer stage), for which the importance of the features also differs. To get a global picture of the important features, we therefore evaluate how often certain features are selected by the classifiers using 100 random train test partitionings with 80 % of the data for training and 20 % of the data for testing. To visualize high confidence relationships between features, we create a graph which has a node for every chosen feature in any of the

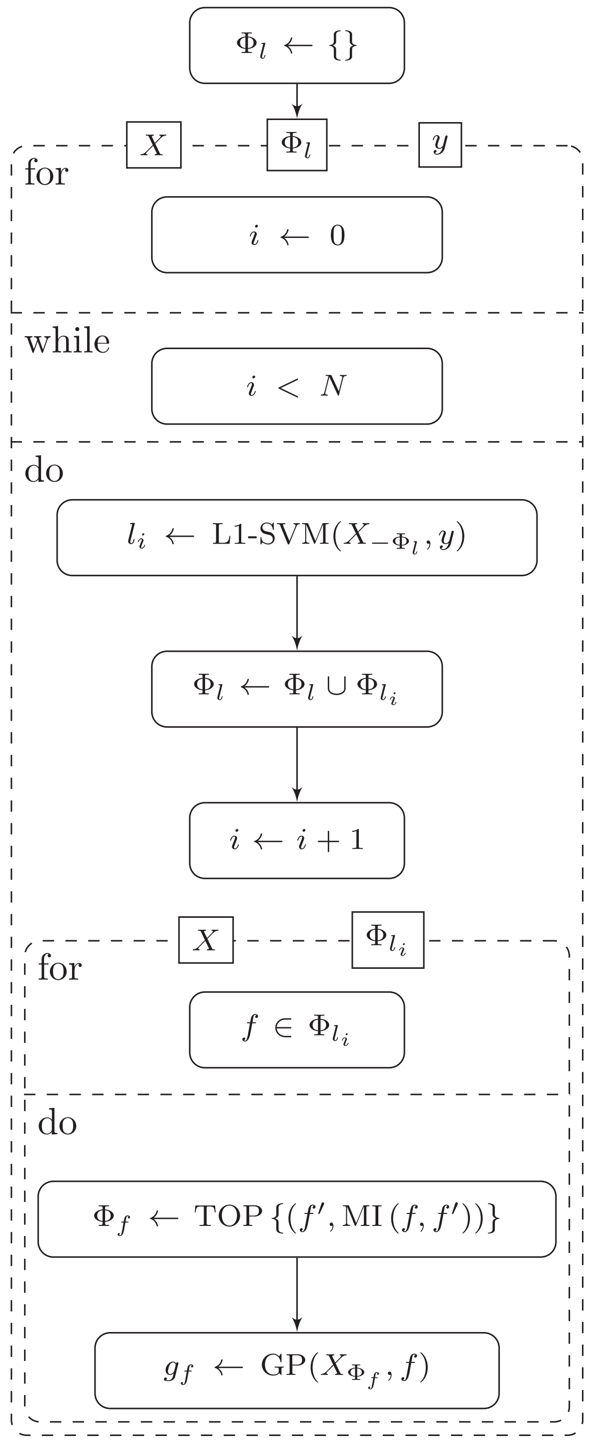
$\Phi_l \leftarrow \{\}$
X
Φ_l
y
for
$i \leftarrow 0$
while
$i < N$
do
$l_i \leftarrow$ L1-SVM$(X_{-\Phi_l}, y)$
$\Phi_l \leftarrow \Phi_l \cup \Phi_{l_i}$
$i \leftarrow i + 1$
X
Φ_{l_i}
for
$f \in \Phi_{l_i}$
do
$\Phi_f \leftarrow$ TOP$\{(f', \text{MI}(f, f'))\}$
$g_f \leftarrow$ GP(X_{Φ_f}, f)

Fig. 1 UML activity diagram of the training process

100 train partitions in any of the individual classifiers. The weight of an edge (s, t) is defined as the number of times the respective features have occurred together in an individual classifier. Then, all edges with low weights are discarded. In order to find a threshold to prune edges according to their weights, a Gaussian kernel density estimate is fit to the weights of the edges, and the threshold is chosen at the 90th percentile. Nodes that have an appearance frequency higher than the threshold are labeled by their gene names and edges having a higher weight than the threshold are kept in the graph.

For illustration purposes, choosing the regularization parameter is done in a way to maximize the number of genes selected with high confidence, as well as minimizing the number of genes pruned out in the process. It is important to remember that considering the results of the method under different regularization parameters is essential to make sure the selected genes possess a high confidence and are also stable regardless of sampling of the training data set.

Implementation details

To compare the performance of our method with other methods, the implementations present in Python *scikit-learn(0.14)* package are taken. In the case of stochastic gradient boosting, the representing class is *GradientBoostingClassifier*, the number of classifiers is set to 100, and to make it sparse and prevent over-fitting, the maximum number of features for splits in trees is set to 5, and the maximum number of layers is set to 2. For AdaBoost, *AdaBoostClassifier* is used, which is an implementation of AdaBoost-SAMME [28], with weak learner set to *DecisionTreeClassifier* with maximum depth set to 2, and the number of weak classifiers set to 100. Parameters of the two boosting algorithms are chosen by a grid search on their parameter space over all the data sets and selecting the parameter sets which give a robust and stable result over all experiments.

As an SVM, ν-SVM with $\nu = 0.25$ is used, once with a linear kernel, and once with an RBF kernel; γ parameter of the RBF kernel is set to $(\text{num of features})^{-1}$. The ν parameter is set to the maximum value for which the optimization function is solvable with *libsvm* for all analyzed data sets [29]. Smaller values cause the SVM to overfit to the data and not generalize well. The Gaussian process's correlation function is a squared-exponential, and MIC is estimated using *minepy* package [30].

The PPI network used in our analysis is from the Human Protein Reference Database (HPRD) [31]. Almost all edges and relationships between proteins that are added to this database are manually extracted from literature by biologists, hence it has a lower rate of edges included in the database for which there is no evidence in the literature.

Results and discussion

Performance comparison

The performance of the method was compared with that of two ensemble methods, AdaBoost and stochastic gradient boosting, as well as an SVM with linear kernel, and an SVM with an RBF kernel. We also included our implementation of the NICK method [4]. We randomly partitioned the data into training and test sets with 80 % of the data for training and 20 % of the data for testing. To compare the performance of the different methods, Area Under the receiver operating characteristic Curve (AUC) [32] was calculated on the test set over the decision values returned by the methods on the individual samples. The process was repeated 100 times to reduce random effects. As seen in Fig. 2, overall performances of all methods are comparable. In some cases a single SVM works better, in some other cases ensemble algorithms give a better performance. However, in most cases an improvement in performance is observed by adding individual learners to the model, with the greatest gains due to the first few individual learners added to the model. In two cases, TCGA-LAML/Vital status and TCGA-LAML/Risk Group, our reported performance measures are significantly lower than other methods. This, however, comes from the fact that we have enforced extreme sparsity measures. The performance of the method increases and reaches the other methods' performance levels if this constraint is relaxed, as reported in Additional file 2. We enforced those sparsity measures for all models to avoid over-fitting. Optimizing the sparsity constraint via cross-validation would have been computationally expensive, which is why we preferred to be conservative. Had we optimized the sparsity constraint, we would have still been able to find the significant features while having similar performance as the other methods. We would like to note that as shown in Additional file 2, for TCGA-LAML/Vital status and TCGA-LAML/Risk Group, the performance of a single learner seems to be better than having multiple learners. This could be due to the fact that the hidden underlying data distribution is more homogeneous than in the other prediction tasks (e.g., there is only one batch). Furthermore, there is generally no free lunch in optimization [33], meaning loosely speaking that there will always be a data set where a novel method performs worse than other methods. We plan to investigate these cases further (i.e., can we estimate when it is better to use one learner instead of multiple learners), and improve our method to tackle the peculiarities of these data sets.

Interpretability of predictions

Here we present the results of running the method on the TCGA-LAML gene expression data set.

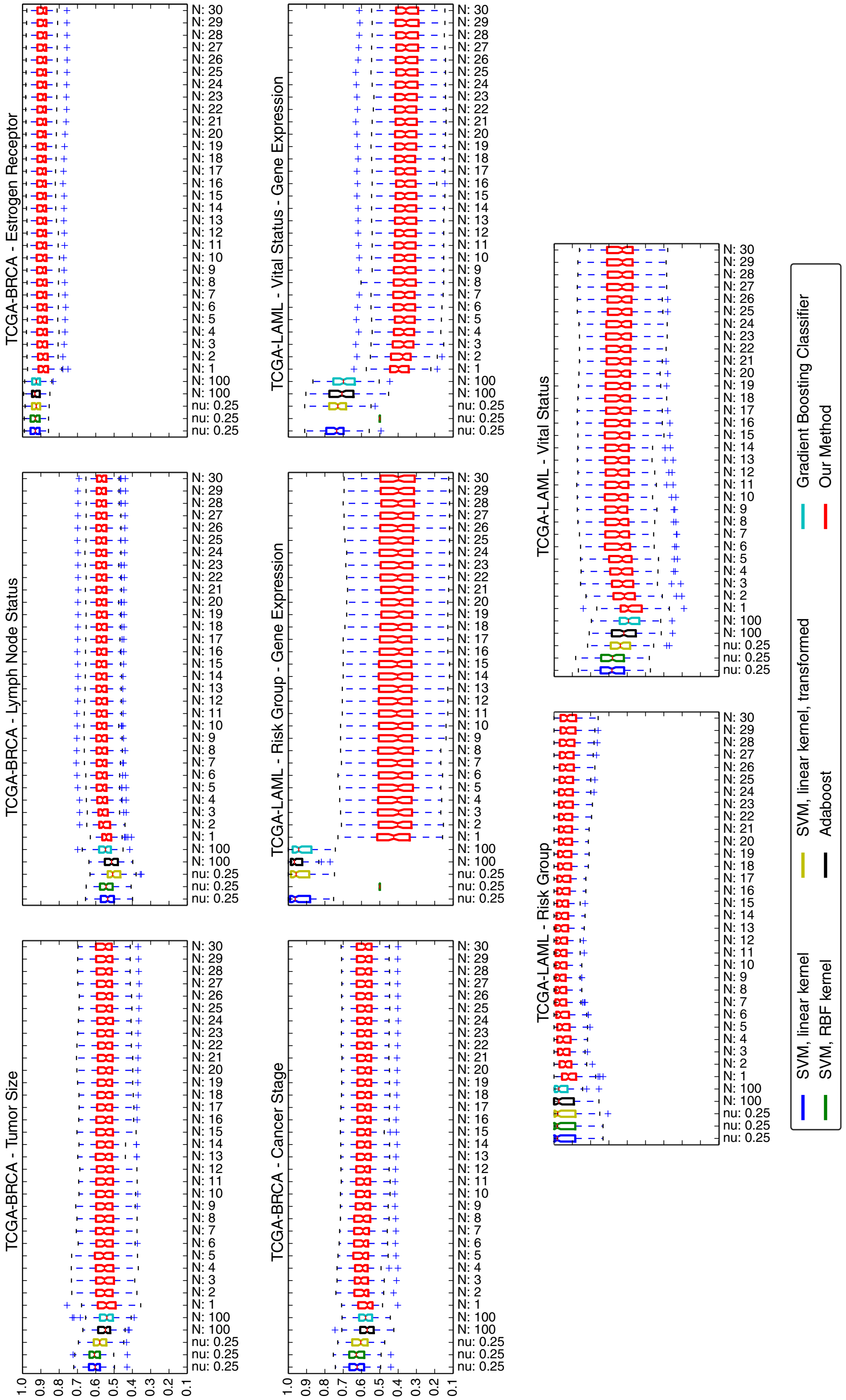

Fig. 2 Performance Summary (AUC). Each box shows a 25–75 % interval, as well as the median, which is shown as a horizontal line in each box

Visualization of features important for a particular test sample

Having a model trained on the data, and given a test sample, it is possible to infer and visualize which individual classifier(s) is (are) influencing the prediction most. To this end, individual learners as well as the features they use are visualized as in Fig. 3a. In this figure, nodes with labels starting with "*L_*" represent individual classifiers, and other nodes are labeled with their respective gene name. The color of the node shows its confidence compared to other nodes; the darker the node, the higher the confidence. In the case of a gene, it is the confidence or reliability of the feature (c_f), and in the case of an individual classifier, it is the overall estimated confidence (c_{l_i}). Edges show which classifier is using which genes in its decision function. The shape of a node represents the individual classifier they belong to.

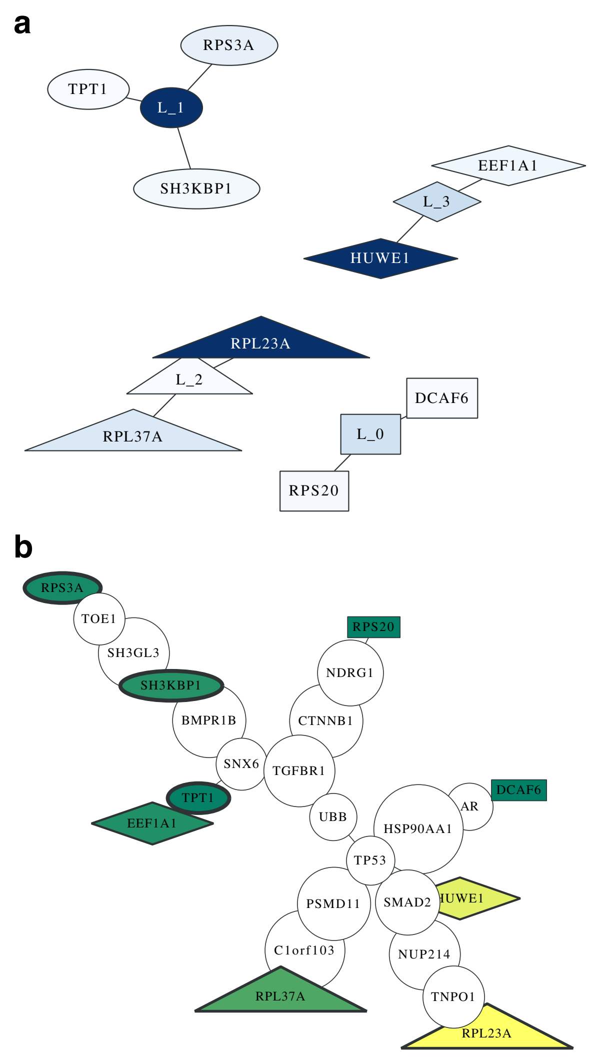

Fig. 3 Visualization of one model. A sample model for TCGA-LAML gene expression data (**a**) individual classifiers and their selected features; higher confidence of a node is shown by a *darker color*, (**b**) selected genes plotted over the PPI network; *green* and *yellow* show low and high confidence respectively, and the thickness of the border of the node shows the respective confidence of the individual classifier to which it belongs

To get a better overview of the individual features that were chosen by the classifiers for the particular test sample, we visualized the corresponding genes on a graph containing information about the PPI network in Fig. 3b. We extracted the PPI information from HPRD as explained before. This way, it is possible to find over- or under-regulated pathways that might be responsible for the label (e.g., cancer stage) of the test sample. Since PPI networks can be quite dense, we removed parts of the induced network. For this purpose we computed each shortest path between all pairs of selected features. Then, the minimum spanning tree of that section was plotted, after removing branches with no selected feature.

Most of the features chosen by any of the classifiers (colored nodes) are not connected to any other chosen feature. It is known that there is in many cases a correlation between expression value of the genes whose corresponding proteins interact [34]. Therefore, a regularized model will only choose a subset of the correlated features. This explains the observation that features selected by a single model can be distant from each other on a PPI network; but if multiple disjoint sparse models are fit to the data, their selected features might happen to be close to each other on the PPI network (e.g., node TPT1 and node EEF1A1 in Fig. 3b).

Visualization of important global features

As explained in Section "Visualization of model predictions", a graph is created from model structures of all 100 random training partitions, and then it is pruned to keep only high confidence nodes and edges. The density estimation of the graph edge weights and the pruned graph are plotted in Fig. 4a, b where the nodes with labels are the ones that are not pruned. The nodes in this figure that do not have any label, are the ones with frequency lower than the corresponding threshold. Among the features considered to be important were features that had previously been linked to leukemia such as SH3KBP1 [35].

What was more intriguing to see was that four out of the seven important features of the TCGA-LAML gene expression data set contained ribosomal proteins when using the risk group label, i.e. RPL37A, RPS20, RPS3A, and RPL23A. For a long time ribosomes were just considered machines that perform an unbiased translation of genes from mRNA to amino acid sequences, but this view has recently been challenged [36]. One new hypothesis is that the ribosome introduces an additional regulatory layer. Therefore, it could very well be that mutations in ribosomal proteins can lead to a misregulation of

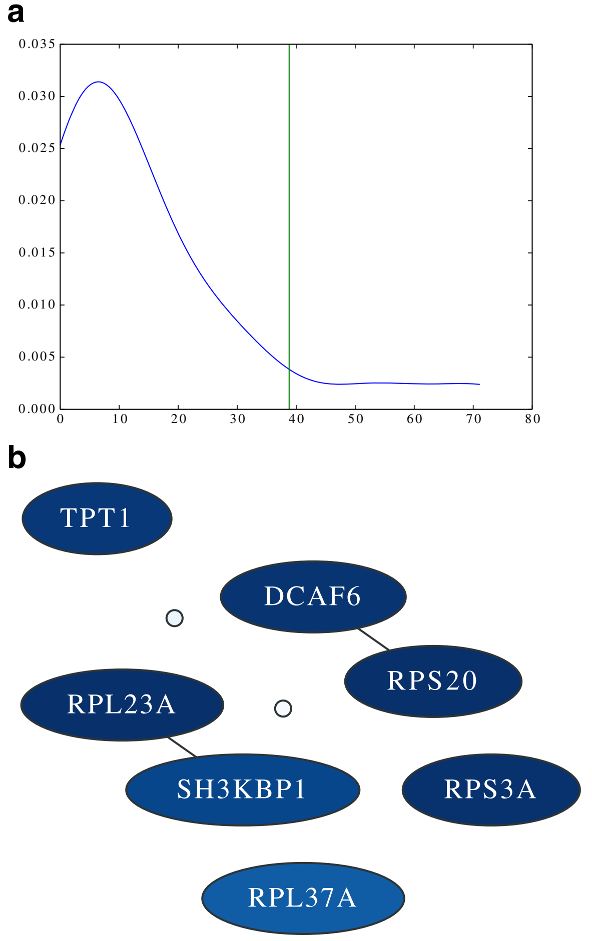

Fig. 4 a Determine pruning threshold. Threshold is determined by finding the point after which, 90 % of the area under the curve is observed from *left* to *right*. The horizontal axis shows the observed frequency or weight of the edges. **b** Important Global Features High confidence nodes and edges of the graph generated from the model on TCGA-LAML gene expression data. Darker color represents higher rate of being selected by a classifier

expression levels of important genes and ultimately to the development of cancer (in this case leukemia). One of the ribosomal proteins we found was RPL23A. It has been shown that loss of RPL23A can impede growth and lead to morphological abnormalities in Arabidopsis Thaliana [36]. Therefore, a mutation in RPL23A might also have severe effects in humans. A missense mutation in RPL23A was recently found in patients having Diamond-Blackfan anemia, which is an inherited form of pure red cell aplasia (related to leukemia) [37]. Note that the model for LAML has low performance for the regularization value chosen. Nevertheless, the features shown here are also the ones with the highest confidence for models learnt with less regularization (with several other additional features). The models with less regularization show similar performance to the other methods shown in Fig. 2.

Conclusions

Machine learning has become more and more popular in many real world scenarios for making sense of large collections of facts. Differences between the data used for training the method and new data for which the label should be predicted can limit the performance of prediction methods on those data. In this work we introduced a method that estimates these potential partial biases and incorporates them into the prediction function. We applied it to gene expression and DNA methylation measurements from cancer patients. Our method has state-of-the-art performance on many different prediction tasks. Furthermore, we show how to make sense of the predictions. Visualizing the important genes can lead to new biological insights, as shown for the TCGA-LAML data set with the risk group label. Instead of mapping the genes to PPI networks, one could also think of mapping them to signaling pathways [38].

Recently, a study showed that most published signatures are not significantly more associated with cancer outcome than random signatures [39]. One of the reasons for this finding is that the data comes from slightly different underlying hidden data distributions. Since our new method estimates this bias and corrects for it by up-weighting the classifiers that have higher confidence, we expect that it should be less susceptible to such differences in the data.

In this work we designed and developed a method that besides being a predictive model, it can be used for two different purposes. It can be used as an exploratory method to reveal potential features used in future studies; and it can be used to different underlying causes of the same disease and with its interpretability help oncologists to choose the treatment accordingly.

We would like to point out that the applicability of our method is not limited to cancer outcome prediction, and it can apply to many more scenarios. The method assumes that the data has enough features to select from, and that there are related features to those selected ones that can be used to estimate their reliability. These are conditions that almost all biological data satisfy, hence the method can be applied to them.

The method also works as a skeleton whose components can be easily substituted. For example, by changing the classifier used in individual learners to a multi-class classifier, the method would work on multi-class problems. For the sake of simplicity and without loss of generality we performed the evaluations only on binary classification problems. Also, due to the structure of our model, one possible approach would be to use a method such as iRDA and use those gene sets as features of individual learners. Whether this approach leads to better results or not requires further research. Also, the combination of maximal information coefficient and Gaussian processes is not the only feasible option, and they can be replaced with other faster methods if the time complexity of the method is of any concern. Some of these alternatives are already available on the *github* repository of the method.

Additional files

Additional file 1: NICK method analysis. (PDF 676 kb)

Additional file 2: Detailed performance measures. (PDF 578 kb)

Acknowledgements
We thank Thomas Lengauer, Sarvesh Nikumbh, and Mathieu Flinders for helpful discussions.

Authors' contributions
AJ: Idea, implementation, testing, interpretation, and writing. NP: Idea, interpretation, and writing. Both authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 23 July 2015 Accepted: 22 April 2016
Published online: 19 July 2016

References

1. Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet. 2013;14(9):618–30. doi:10.1038/nrg3542.
2. Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics. 2005;21(2):171–8. doi:10.1093/bioinformatics/bth469.
3. Chuang H-Y, Lee E, Liu Y-T, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol. 2007;3:140. doi:10.1038/msb4100180.
4. Lavi O, Dror G, Shamir R. Network-induced classification kernels for gene expression profile analysis. J Comput Biol. 2012;19(6):694–709. doi:10.1089/cmb.2012.0065.
5. Bradley PS, Mangasarian OL. Feature selection via concave minimization and support vector machines. In: Proceedings of the International Conference on Machine Learning. San Francisco, California: Morgan Kaufmann Publishers; 1998. p. 82–90.
6. Schapire RE, Singer Y. Improved boosting algorithms using confidence-rated predictions. Mach Learn. 1999;37(3):297–336.
7. Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE. Adaptive mixtures of local experts. Neural Comput. 1991;3(1):79–87.
8. Bishop CM, Svenskn M. Bayesian hierarchical mixtures of experts. In: Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence. San Francisco, California: Morgan Kaufmann Publishers Inc; 2002. p. 57–64.
9. Saeys Y, Abeel T, Van de Peer Y. Robust feature selection using ensemble feature selection techniques. In: Machine Learning and Knowledge Discovery in Databases. Heidelberg, Germany: Springer; 2008. p. 313–25.
10. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn. 2002;46(1–3): 389–422.
11. Lai H-M, Albrecht AA, Steinhöfel KK. irda: a new filter towards predictive, stable, and enriched candidate genes. BMC Genomics. 2015;16(1):1.
12. Shi J, Walker MG. Gene set enrichment analysis (gsea) for interpreting gene expression profiles. Curr Bioinformatics. 2007;2(2):133–7.
13. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Nat Acad Sci USA. 2005;102(43): 15545–50.
14. Nam D, Kim SY. Gene-set approach for expression pattern analysis. Brief Bioinformatics. 2008;9(3):189–97.
15. Dopazo J. Functional interpretation of microarray experiments. Omics: J Integr Biol. 2006;10(3):398–410.
16. The Cancer Genome Atlas Network. The Cancer Genome Atlas (TCGA). 2006. https://tcga-data.nci.nih.gov/tcga/. Accessed 2013.
17. The Cancer Genome Atlas Network. Genomic and epigenomic landscapes of adult *de novo* acute myeloid leukemia. N Engl J Med. 2013;368(22):2059–74.
18. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61–70.
19. Knight WA, Livingston RB, Gregory EJ, McGuire WL. Estrogen receptor as an independent prognostic factor for early recurrence in breast cancer. Cancer Res. 1977;37(12):4669–71.
20. Goldhirsch A, Glick JH, Gelber RD, Coates AS, Senn HJ. Meeting highlights: international consensus panel on the treatment of primary breast cancer. J Clinical Oncol. 2001;19(18):3817–27.
21. National Institutes of Health Consensus Development Panel and others. National institutes of health consensus development conference statement: adjuvant therapy for breast cancer, november 1—3, 2000. J Natl Cancer Inst. 2001;93(13):979–89.
22. Dietterich T. Ensemble Learning. The Handbook of Brain Theory and Neural Networks. Second Edition. Cambridge, MA: The MIT Press; 2002. p. 405–8.
23. Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal. 2002;38(4):367–78.
24. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput System Sci. 1997;55(1): 119–39.
25. Heppner GH. Tumor heterogeneity. Cancer Res. 1984;44(6):2259–65.
26. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, Lander ES, Mitzenmacher M, Sabeti PC. Detecting novel associations in large data sets. Science. 2011;334(6062):1518–24.
27. Lin H-T, Lin C-J, Weng RC. A note on Platt's probabilistic outputs for support vector machines. Mach Learn. 2007;68(3):267–76.
28. Zhu J, Zou H, Rosset S, Hastie T. Multi-class AdaBoost. Stat. Interface. 2009;2(3):349–60.
29. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol. 2011;2:27–12727.
30. Albanese D, Filosi M, Visintainer R, Riccadonna S, Jurman G, Furlanello C. minerva and minepy: a c engine for the mine suite and its r, python and matlab wrappers. Bioinformatics. 2013;29(3):407–8.
31. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi T, Gronborg M, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res. 2003;13(10): 2363–71.
32. Egan JP. Signal detection theory and ROC analysis. New York: Academic Press; 1975.
33. Wolpert DH, Macready WG. No free lunch theorems for optimization. Evol Comput IEEE Trans. 1997;1(1):67–82.
34. Jansen R, Greenbaum D, Gerstein M. Relating whole-genome expression data with protein-protein interactions. Genome Res. 2002;12(1):37–46.
35. Adélaïde J, Gelsi-Boyer V, Rocquain J, Carbuccia N, Birnbaum DJ, Finetti P, Bertucci F, Mozziconacci MJ, Vey N, Birnbaum D, Chaffanet M. Gain of CBL-interacting protein, a possible alternative to CBL mutations in myeloid malignancies. Leukemia. 2010;24(8):1539–41. doi:10.1038/leu.2010.135.
36. Xue S, Barna M. Specialized ribosomes: a new frontier in gene regulation and organismal biology. Nat Rev Mol Cell Biol. 2012;13(6):355–69. doi:10.1038/nrm3359.
37. Gazda HT, Preti M, Sheen MR, O'Donohue MF, Vlachos A, Davies SM, et al. Frameshift mutation in p53 regulator RPL26 is associated with multiple physical abnormalities and a specific pre-ribosomal RNA processing defect in diamond-blackfan anemia. Hum Mutat. 2012;33(7): 1037–44. doi:10.1002/humu.22081.
38. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 2014;42(Database issue):199–205. doi:10.1093/nar/gkt1076.
39. Venet D, Dumont JE, Detours V. Most random gene expression signatures are significantly associated with breast cancer outcome. PLoS Comput Biol. 2011;7(10):1002240. doi:10.1371/journal.pcbi.1002240.