Bioinformatics Advance Access published November 25, 2009

Robust biomarker identification for cancer diagnosis with ensemble feature selection methods

Thomas Abeel[1,2], Thibault Helleputte[3,4], Yves Van de Peer[1,2], Pierre Dupont[3,4], Yvan Saeys[1,2,*]

[1]Department of Plant Systems Biology, VIB, Technologiepark 927, 9052 Gent, Belgium.
[2]Department of Molecular Genetics, Ghent University, Gent, Belgium.
[3]Department of Computing Science and Engineering INGI, Université catholique de Louvain, Belgium.
[4]Machine Learning Group, Université catholique de Louvain, Belgium.

Associate Editor: Prof. Thomas Lengauer

ABSTRACT

Motivation: Biomarker discovery is an important topic in biomedical applications of computational biology, including applications such as gene and SNP selection from high dimensional data. Surprisingly, the stability with respect to sampling variation or robustness of such selection processes has received attention only recently. However, robustness of biomarkers is an important issue, as it may greatly influence subsequent biological validations. In addition, a more robust set of markers may strengthen the confidence of an expert in the results of a selection method.

Results: Our first contribution is a general framework for the analysis of the robustness of a biomarker selection algorithm. Secondly, we conducted a large-scale analysis of the recently introduced concept of ensemble feature selection, where multiple feature selections are combined in order to increase the robustness of the final set of selected features. We focus on selection methods that are embedded in the estimation of support vector machines (SVMs). SVMs are powerful classification models that have shown state-of-the-art performance on several diagnosis and prognosis tasks on biological data. Their feature selection extensions also offered good results for gene selection tasks. We show that the robustness of SVMs for biomarker discovery can be substantially increased by using ensemble feature selection techniques, while at the same time improving upon classification performances. The proposed methodology is evaluated on four microarray data sets showing increases of up to almost 30% in robustness of the selected biomarkers, along with an improvement of about 15% in classification performance. The stability improvement with ensemble methods is particularly noticeable for small signature sizes (a few tens of genes), which is most relevant for the design of a diagnosis or prognosis model from a gene signature.

Contact: yvan.saeys@psb.ugent.be

*to whom correspondence should be addressed

1 INTRODUCTION

The identification of substances that are indicative of a specific biological state, broadly referred to as biomarkers nowadays, is an important research topic in the biomedical field. Especially in cancer diagnostics, technologies such as microarrays, and more recently also mass spectrometry, have become an established technique to compare diseased samples to control samples. From a machine learning point of view, the selection of biomarkers in this context can be stated as a feature selection problem for a classification task, where the aim is to find a small set of features (markers) that best explains the difference between the disease and the control samples.

Feature selection offers a number of advantages, including more powerful classification models by eliminating irrelevant or noisy features (Krishnapuram *et al.*, 2004), more compact and faster models by constructing them using only a small subset of the original set of features, and the ability to focus on a subset of relevant features, which can be used for the discovery of new knowledge (Guyon and Elisseeff, 2003). Future clinical tests can then potentially be built at a cheaper cost on fewer markers.

Feature selection techniques can be broadly characterized into three classes, depending on how they interact with the estimation of the classification model (Saeys *et al.*, 2007). Filter methods work independently of the classifier design, and perform feature selection by looking at the intrinsic properties of the data. In contrast, wrapper and embedded methods perform feature selection by making use of a specific classification model. While wrapper methods employ a search strategy in the space of possible feature subsets, guided by the predictive performance of a classification model, embedded methods make use of the classification model internal parameters to perform feature selection. Embedded methods show a better computational complexity than wrapper methods, especially in high-dimensional spaces.

This work focuses on the use of feature selection techniques for biomarker discovery from microarray data. It is common practice for a domain expert to start validating the biomarkers selected by the feature selection algorithm in a top-down fashion, with techniques such as RT-PCR. However, different feature selection techniques may result in different rankings of the features. The same

feature selection technique may produce drastically different results depending on the chosen setting of the parameters of the method. To make matters even worse, many of the current data sets are described by a number of features that generally exceed the number of available training samples by orders of magnitude. These so-called *wide data* may lead to even more variation in the final ranking of the features. In the particular context of genomic data, a stable feature selection technique is desirable. Selection of relevant genes for a given pathology on different sub-samplings of the patients should produce nearly the same results since the biological process generating the data is assumed to be largely common for all patients, at least without confounding factors.

Surprisingly, the analysis of the stability or robustness of biomarker selection techniques is only a topic of recent interest, and has not yet made it into the mainstream methodology for biomarker discovery. Therefore, we propose a general experimental setup for stability analysis that can be easily included in any biomarker identification pipeline. In addition, we also present new techniques to increase the stability of the final set of selected biomarkers, using the recently introduced concept of ensemble feature selection (Saeys *et al.*, 2008). Ensemble methods have originally been developed to enhance classification performance (Dietterich, 2000). The general idea of that family of techniques consists in combining lots of different models in a global, more robust, model. The different models are typically built on different sub-samplings of the original data set. Ensemble method ideas have recently been ported to *wide data* (Long and Berlian Vega, 2003; Dettling, 2004) and in particular to feature selection (Saeys *et al.*, 2008).

In the present study we use support vector machines (SVMs) classifiers (Boser *et al.*, 1992). These classifiers are interesting because the number of parameters to be estimated essentially depends on the number of samples rather than the number of features, which is particularly relevant with very small sample-to-feature ratios. Such models also offer state-of-the-art classification performance on a wide range of applications (Schoelkopf and Smola, 2002). Moreover, SVMs have been extended to form embedded feature selection methods. A prominent approach among them is the RFE selection method (Guyon *et al.*, 2002). A linear SVM is a classification model for which the influence of each dimension, here a specific gene, is explicitly available. RFE precisely uses this property to remove the least important features and iteratively re-estimates a classifier on the remaining features.

Our experiments on four cancer diagnosis microarray data sets show that the robustness of gene selection methods can be significantly improved by extending them with an ensemble procedure. The relative performances of various ways of combining the individual signatures is also rigorously assessed.

2 MATERIALS AND METHODS

Our experimental data consists of 4 cancer diagnosis microarrays data sets which are first described. Next we propose a general evaluation protocol to evaluate both the stability and the classification performance of gene selection methods. We discuss data normalization and present our reference gene selection method, known as RFE (Guyon *et al.*, 2002). Much simpler selection methods do exist (Guyon and Elisseeff, 2003) but RFE is chosen here as a baseline because it is known to provide state-of-the-art classification performance and was originally applied precisely in the context of gene selection for cancer classification. Besides, like most feature selection methods embedded with a classifier estimation, RFE is intrinsically multivariate in the sense that it evaluates the relevance of several features considered jointly. In contrast, a univariate method evaluates the relevance of each feature individually. The latter is often simpler computationally but the former is more refined from a data analysis viewpoint and also biologically more relevant, because genes are known to interact in many ways and are often co-regulated.

In the following, we introduce the ensemble feature selection approach which relies on different sub-samplings of the original data to build different signatures. We detail two different aggregation methods to build a consensus from the various signatures.

Table 1. Overview of the data sets used. SDR refers to the ratio between the number of samples and the number of dimensions (or features).

Name	Ref.	# samples (+/-)	# dim.	SDR
Leukemia	(Golub *et al.*, 1999)	72 (47/25)	7,129	0.010
Colon	(Alon *et al.*, 1999)	62 (40/22)	2,000	0.031
Lymphoma	(Alizadeh *et al.*, 2000)	45 (22/23)	4,026	0.011
Prostate	(Singh *et al.*, 2002)	102 (52/50)	6,033	0.017

2.1 Microarray data sets

The datasets we use in this work are microarray datasets. Their main characteristics are summarized in Table 1. They share common characteristics such as a very low samples/dimensions ratio.

The original studies on these data sets stress the limitations of univariate methods, which look at the influence of each gene individually, and the issue of signature robustness. They trigger the interest of considering a multivariate technique, like RFE, and of trying to improve its robustness.

The Leukemia dataset was produced in a study aimed at building a model to discriminate between Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL) tissues (Golub *et al.*, 1999). It is the largest dataset used in this paper in terms of number of features (7,129). A set of 50 genes was selected in a univariate way according to a correlation criterion with the class label. Those genes where combined afterwards to build a simple weighted vote classifier. The model was tested on an independent test set of 34 Leukemia samples, making correct predictions for 29 among them. Authors mention the difficulty of choosing the right set of informative genes, given that lots of them were highly correlated with the ALL-AML distinction, and stress the need of a stable selection approach.

The Colon Cancer dataset is made of samples from 40 tumor and 22 normal colon tissues probed by an Affymetrix microarray chip measuring more than 6,500 genes (Alon *et al.*, 1999). The data set was, however, published after a pre-filtering step and the resulting samples include only 2,000 genes. The task described in the original study is to determine if groups of patients could automatically be constructed by a clustering algorithm. Clustering separated cancerous from noncancerous tissue and cell lines from in vivo tissues on the basis of subtle distributed patterns of genes even when expression of individual genes varied only slightly between the tissues. Since annotations of the samples are available, we address here the binary classification problem of predicting whether a sample corresponds to a tumor versus a normal colon tissue.

The Lymphoma dataset comes from a study on Diffuse Large B-Cell Lymphoma (Alizadeh *et al.*, 2000). The task is here to discriminate between two types of Lymphoma based on gene expression measured by microarray technology. The authors originally performed a clustering analysis based on similarity measure related to correlation. This dataset has expression measurements for 4,026 genes.

The Prostate dataset was first published in (Singh *et al.*, 2002). One of the tasks addressed by the authors is to build a model able to discriminate between normal and tumor prostate tissue, based on microarray data. A univariate selection criterion (Golub's Signal-to-Noise Ratio) is applied on each gene and its significance is assessed by a permutation test. The

most significant genes are selected and a predictive model is built on this signature with a k-nearest neighbor classifier. The number of genes found to be significantly differentially expressed in the two classes according to the Golub's ratio is higher than 400. This result shows the difficulty of identifying small and stable signatures. Each selected gene has been individually compared with clinical factors but no significant correlation was found.

2.2 Biomarker evaluation protocol

In order to analyze the stability of a biomarker selection algorithm, we propose to generate slight variations of the original data set, and compare the outcome of the marker selection algorithm across these different variations. The rationale behind this is that, for a stable marker selection algorithm, small changes in the training set should not yield big changes in the set of finally selected markers. This is consistent with what a domain expert would expect from a marker selection algorithm: adding or deleting a few samples should not drastically modify the top-ranked markers identified by the algorithm.

To implement the strategy of generating slight variations of the data set, a sub-sampling approach is proposed: a large number (e.g. 500) of data sets can be generated by subsampling the original data set without replacement. As microarray datasets generally contain few samples, we suggest to generate subsamplings containing 90% of the samples of the original data set.

We argue that stability matters, but stability alone is not a good quality measure. Indeed, one could conceive a trivial selection algorithm which would always return the same features no matter which samples it receives as input. The resulting set of features would be perfectly stable but likely irrelevant for the classification task. Hence, stability needs to be assessed together with classification performance. We propose to use the same subsamplings - each containing 90% of the original data set - as training sets to select features and estimate the performance of a classifier. The remaining 10% of the data can be used each time as an independent validation set to evaluate classification performance. Since we are considering typically 500 independent partitionings into 90% training and 10% validation, we reduce the risk of over-optimistic results of traditional cross-validation experiments on small sample domains (Braga-Neto and Dougherty, 2004).

In the following paragraphs, we detail the exact performance metrics used to assess the stability of biomarker lists and the average classification performances.

2.2.1 Stability measure Let us first formalize the experimental setup as follows. We consider a data set $\mathcal{X} = \{x_1, \ldots, x_M\}$ with M instances and N features. Then, k subsamplings of size $\lceil xM \rceil$ $(0 < x < 1)$ are drawn randomly from $\mathcal{X}$, where in our experiments $k = 500$ and $x = 0.9$. Subsequently, feature selection is performed on each of the k subsamplings, and a marker set -further referred to as a *signature*- of a given size is selected.

Here, following (Kalousis *et al.*, 2007), we take a similarity-based approach where feature stability is measured by comparing the signatures selected on each of the k subsamplings. The more similar all signatures are, the higher the stability measure will be. The overall stability S_{tot} can then be defined as the average over all pairwise similarity comparisons between all signatures on the k subsamplings:

$$S_{\text{tot}} = \frac{2\sum_{i=1}^{k}\sum_{j=i+1}^{k} \text{KI}(\mathbf{f}_i, \mathbf{f}_j)}{k(k-1)}$$

where $\mathbf{f}_i$ represents the signature obtained by the selection method on subsampling i $(1 \leq i \leq k)$, and $\text{KI}(\mathbf{f}_i, \mathbf{f}_j)$ is the Kuncheva index; a stability index between $\mathbf{f}_i$ and $\mathbf{f}_j$, defined as follows (Kuncheva, 2007):

$$\text{KI}(\mathbf{f}_i, \mathbf{f}_j) = \frac{r \cdot N - s^2}{s \cdot (N - s)} = \frac{r - \frac{s^2}{N}}{s - \frac{s^2}{N}}$$

where $s = |\mathbf{f}_i| = |\mathbf{f}_j|$ denotes the signature size and $r = |\mathbf{f}_i \cap \mathbf{f}_j|$ is the number of common elements in both signatures. The Kuncheva index satisfies $-1 < \text{KI}(\mathbf{f}_i, \mathbf{f}_j) \leq 1$ and the greater its value, the larger the number of commonly selected features in both signatures. The $\frac{s^2}{N}$ term in this index corrects a bias due to the chance of selecting common features among two signatures chosen at random. A negative index reflects that feature sharing is mostly due to chance. This correction term pleads for using the Kuncheva index, instead of other stability indices such as the Jaccard index (Kalousis *et al.*, 2007). In the sequel, the overall stability S_{tot} is simply denoted KI.

2.2.2 Classification performance measure To compare the classification performance of different methods, we use the area under the Receiver Operator Curve (ROC, Provost and Fawcett (1997)), further abbreviated as AUC. This area is defined by a function of sensitivity and specificy, frequently used in clinical settings.

2.3 Data normalization

The objective of data normalization is to enhance similarity of genes sharing a common expression pattern throughout the data, but in different ranges of absolute expression values. We use here an IQR-normalization procedure. The normalized expression value $\bar{f}_{ij}$ is defined as follows.

$$\bar{f}_{ij} = \frac{f_{ij} - m_j}{IQR_j/1.35}$$

where f_{ij} is the original expression value of gene j from sample i, m_j is the median of expression of this gene over all samples, and IQR_j stands for the gene-specific interquartile range (Tukey, 1977). The IQR-normalization is more robust to the presence of outliers than a classical Z-score (centering to the mean with unit standard deviation) but the 1.35 scaling factor makes both normalization equivalent whenever the data happens to be normally distributed. The normalization parameters for each gene are always estimated from the training samples only and applied subsequently to the validation samples.

2.4 Embedded feature selection with SVMs

Our reference classifier is a linear SVM (Boser *et al.*, 1992). SVMs are known to scale well to high-dimensional spaces, and have shown state-of-the-art performance in many problems in computational biology (Ben-Hur *et al.*, 2008). Furthermore, a linear SVM offers the additional advantage that it contains an embedded capability for feature selection. As a linear SVM essentially consists of a separating hyperplane in the input space, the absolute values of the weights of each dimension in the hyperplane can be regarded as the contribution (importance) of each dimension (feature) to the multivariate decision of the hyperplane. As a result, these weights can be used to rank the features from most important to least important, which is the rationale for the recursive feature elimination algorithm (RFE, (Guyon *et al.*, 2002)).

In order to use RFE for feature selection, a recursive procedure is started that adopts a backward elimination strategy to iteratively remove features. Starting from the full feature set, a linear SVM is estimated from the training samples and features are sorted according to the absolute value of their weight in the hyperplane. Subsequently the least important features are eliminated and a linear SVM is re-estimated on the same samples but restricted to the remaining set of features. This process is iterated until all features have been removed or a desired number of features is reached.

An internal parameter of the RFE method is the fraction E of features to eliminate at each step, which greatly influences the computational complexity of the method. Decreasing E increases the computational cost since less features are dropped at each iteration but possibly offers a more refined selection. In our experiments, we chose to drop $E = 20\%$ features at each iteration by default. An additional sensitivity analysis is reported in our experiments to check the precise influence of this parameter. When setting $E = 100\%$, RFE reduces to a single SVM estimation which ranks all the features in one step. When setting $E < 100\%$, the overall ranking of features is constructed iteratively from worst to best features. Whenever features are removed, they are sorted according to their absolute weight value

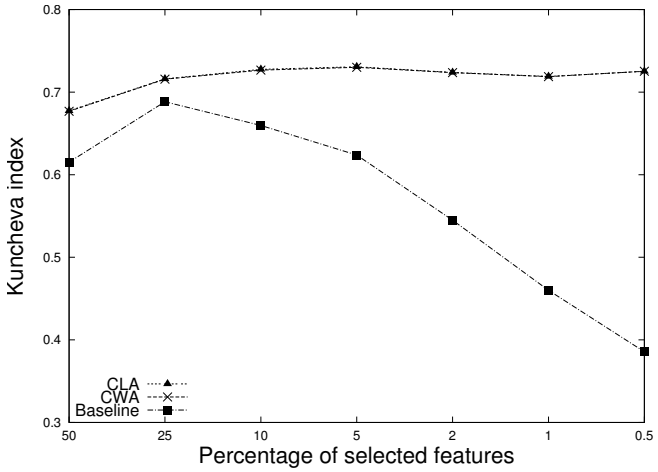

Fig. 1. Stability of the baseline method (original RFE) and the ensemble methods for prostate. We used 40 bootstraps and used RFE with $E = 20\%$.

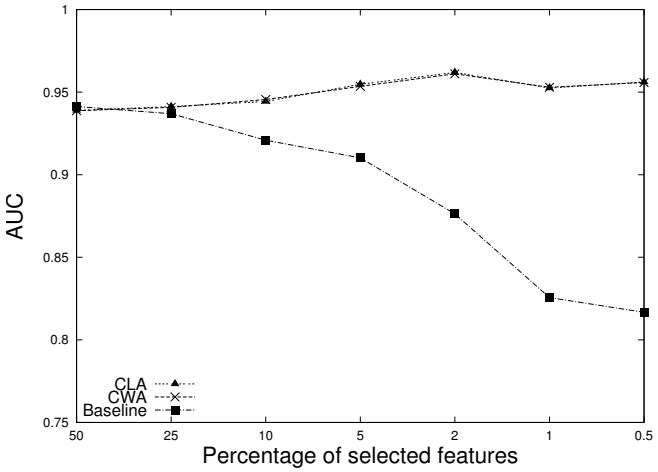

Fig. 2. Classification performances of the baseline method (original RFE) and the ensemble method for prostate. We used 40 bootstraps and used RFE with $E = 20\%$.

in the hyperplane at the iteration of removal and added at the top of the current ranking.

Another internal parameter of the SVM learning procedure is the so-called regularization parameter C. An SVM aims at classifying correctly training samples with a certain margin, which is not always possible when the training samples cannot be perfectly separated in two classes by a hyperplane. The C parameter precisely controls how strong margin errors (samples which cannot be classified with the prescribed margin) are penalized. We use the simple default value $C = 1$ in our reported experiments. We also investigated how to optimize C using a grid search with an internal 5-fold cross-validation on the training part but the additional computational complexity only offered marginal performance gains.

2.5 Ensemble feature selection

To increase the stability of feature selection algorithms, we further elaborated on the recently introduced concept of ensemble feature selection techniques (Saeys *et al.*, 2008). Ensemble feature selection techniques use an idea similar to ensemble learning for classification (Dietterich, 2000): in a first step, a number of different feature selectors are used, and in a final phase the output of these separate selectors is aggregated and returned as the final (ensemble) result. We focus on the analysis of ensemble feature selection techniques using linear SVMs and RFE as the feature selection mechanism.

Starting from a particular training set, i.e. one of the 500 subsamplings containing 90% of the data, our aim is now to generate a diverse set of RFE feature selections. Because the RFE procedure is deterministic, the only chance to generate diversity in the selection is to perform it on different training samples. To this end, we make use of the bootstrapping method, a well-established technique in statistics to reduce variance (Efron, 1979). By drawing (with replacement) different bootstrap samples of the training data, we can apply RFE to each of these bootstrap samples, and thus obtain a diverse set of feature rankings. More formally, we take an ensemble EFS consisting of t feature selectors, EFS $= \{F_1, F_2, \ldots, F_t\}$, then we assume each F_i provides a feature ranking $\mathbf{f}_i = (f_i^1, \ldots, f_i^N)$, where f_i^j denotes the rank of feature j in bootstrap i. The best feature is assigned rank 1, and the worst one rank N.

To aggregate the different rankings, obtained by bootstrapping the training data, into a final signature we propose two aggregation schemes. These aggregation schemes differ in the way the aggregation (sum) of the individual rankings is calculated. A general formulation for the ensemble ranking $\mathbf{f}$, obtained by summing the ranks over all bootstrap samples is as follows:

$$\mathbf{f} = (\sum_{i=1}^{t} w_i \mathbf{f}_i^1, \ldots, \sum_{i=1}^{t} w_i \mathbf{f}_i^N) \quad (1)$$

where w_i denotes a bootstrap dependent weight.

2.5.1 Complete linear aggregation (CLA) This method uses the complete ranking of all the features to create the ensemble result. The ensemble ranking $\mathbf{f}$ is then obtained by just summing the ranks over all bootstrap samples. In the general formulation (1), this amounts to setting all weights w_i equal to 1. To select the final set of features for a signature of size s, the s features with the lowest summed rank are selected from $\mathbf{f}$.

2.5.2 Complete weighted linear aggregation (CWA) This method is a variation on the previous method, where we no longer just sum all scores, but additionally weigh the scores of each bootstrap ranking. The weight assigned to a bootstrap ranking is the AUC obtained by a linear SVM, trained on the bootstrap samples and evaluated on the out-of-bag (OO) samples. In formulation (1), this amounts to setting $w_i = \text{OO-AUC}_i$.

2.6 Implementation and availability

All algorithms for feature selection, classification, as well as the extensions to ensemble feature selection and stability analysis were used as implemented in Java-ML, a publicly available, open source Java machine learning library (`http://java-ml.sf.net/`), or implemented in `R`, a publicly available and open source language for statistical applications (`http://r-project.org`).

3 RESULTS

We report here the experimental evaluations on the four cancer diagnosis microarray data sets considered in the present study. Our reference method serving as a baseline is the RFE approach. It already offers state-of-the-art classification performance and a multivariate selection mechanism for evaluating the combined relevance of sets of markers. Our results show that the stability of biomarker selection with RFE, as well as the classification performance can be significantly improved with the proposed ensemble methodology. The relative performance of the different ways of building a consensus from distinct individual signatures is

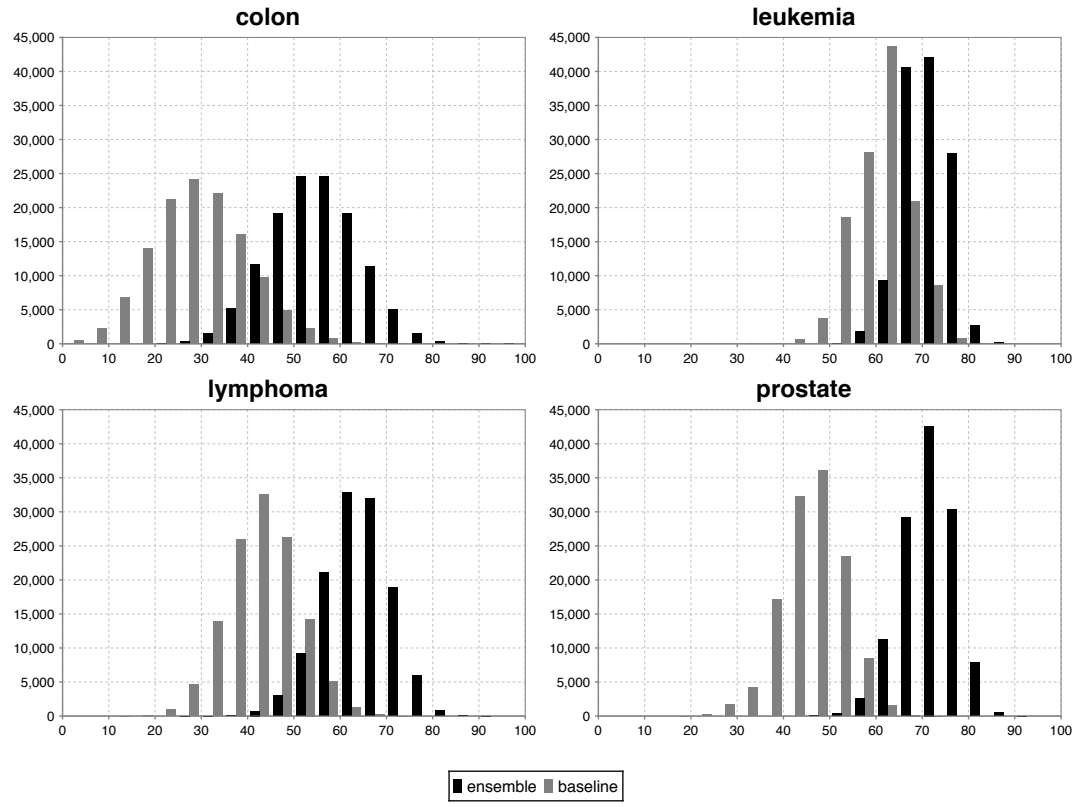

Fig. 3. Distribution plots of the pair wise stabilities for the four data sets. We used 40 bootstraps, eliminated $E = 20\%$ features at each iteration of RFE, used a signature size of 1% and chose the CLA aggregation model. We used fixed-width bins of 5%.

carefully assessed. Results on the variance of the pairwise stability between two signatures are also discussed, as well as the sensitivity to the setting of internal parameters of the marker selection methods.

3.1 Ensemble feature selection methods improves classification performance and biomarker stability

As a first experiment, we compare our newly proposed ensemble RFE method to the traditional RFE setting, analyzing the stability and classification performance for each of the four cancer data sets used in this study. Figures 1 and 2 display the results for the Prostate dataset, using a default configuration where the number of bootstrap rounds to create the ensemble is set to 40, and RFE was applied eliminating $E = 20\%$ of the features at each iteration. Results for the other datasets can be found in Supplementary figures 1 and 2. The default configuration was based on earlier work (Saeys *et al.*, 2008), where it was shown that these parameter settings yielded a good default. Robustness of the selected signatures (marker sets) was measured by the Kuncheva index (KI), and the AUC was used to measure the classification performance.

It can be observed that the ensemble methods CLA and CWA clearly improve upon the baseline, both in terms of stability and classification performance. Moreover, the gains increase as signature sizes get smaller. In three out of four datasets, the ensemble methods even perform better with fewer than with all features (see Supplementary Material), thus showing that ensemble methods are better capable of eliminating noisy and irrelevant dimensions.

In the above analysis we reported average classification and stability performances over 500 distinct subsamplings from each original data set. The proposed stability metric is an average over all pairwise comparisons of two signatures built using different subsamplings. Figure 3 details this analysis by reporting the histogram of stability values across pairwise comparisons. We observed that the ensemble CLA method improves the average stability over the baseline, since there is a systematic shift of the histograms with respect to the baseline, with no influence on the variance of the stability since the respective histograms have the same spread.

3.2 Sensitivity analysis with respect to internal parameters

Up to now, performance metrics were computed using default parameters which were proven useful in our preliminary study. We further assessed the influence of two internal parameters that may influence the results: the number of bootstrap rounds used by the ensemble methods and the fraction of features discarded at each iteration of RFE (baseline or ensemble versions).

3.2.1 Sensitivity with regard to the number of bootstrap samples Figures 4 and 5 show the stability and classification performance

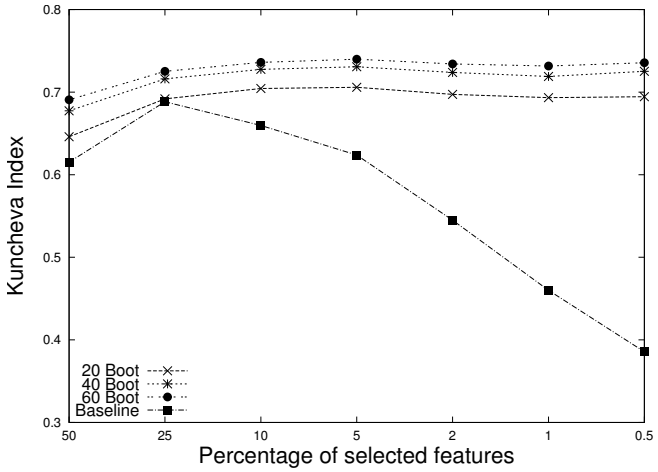

Fig. 4. Stability for several numbers of bootstrap rounds for the construction of an ensemble signature for prostate. We used the CLA aggregation method and eliminated 20% of the features at each iteration of RFE. The baseline is the original RFE on the full training sets without bootstrap.

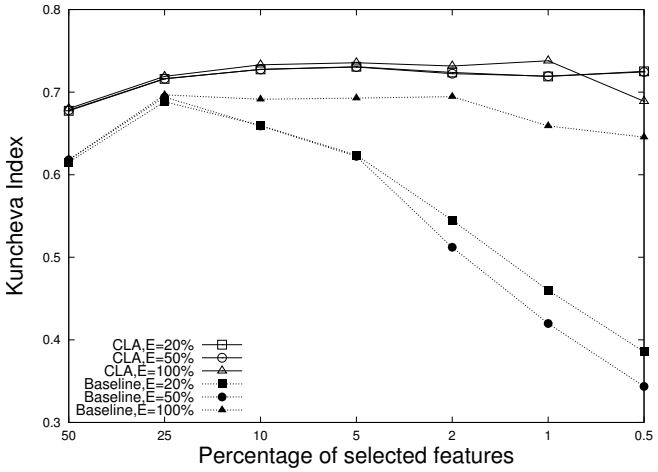

Fig. 6. Stability with regard to a varying number of features to eliminate during RFE. Results represent the CLA aggregation and constructed using 40 bootstrap samples.

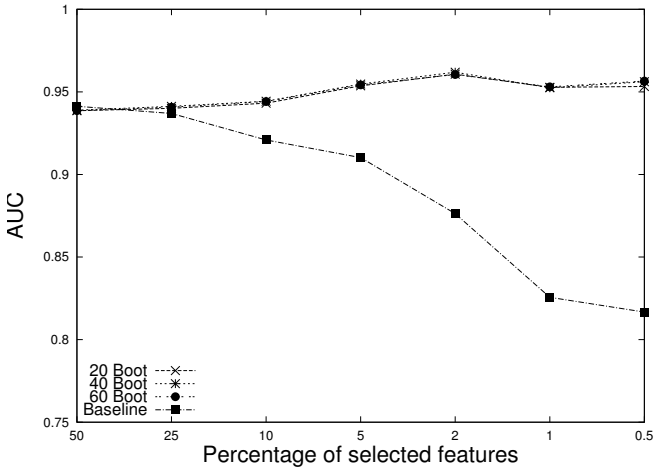

Fig. 5. Classification performances for several numbers of bootstrap rounds for the construction of an ensemble signature for prostate. We used the CLA aggregation method and eliminated 20% of the features at each iteration of RFE. The baseline is the original RFE on the full training sets without bootstrap.

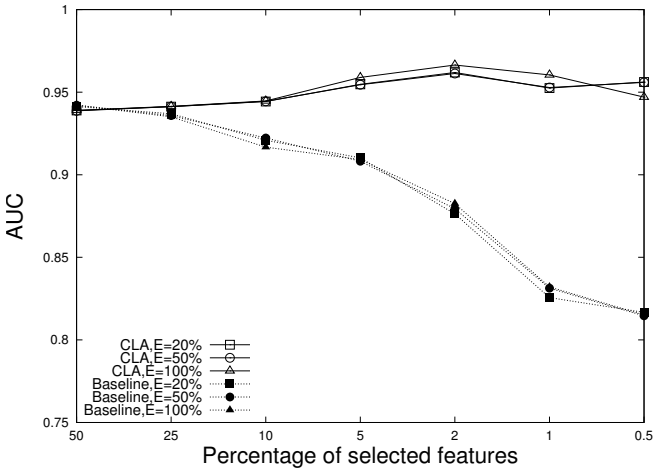

Fig. 7. Classification performance with regard to a varying number of features to eliminate during RFE. Results represent the CLA aggregation method and were constructed using 40 bootstrap samples.

with regard to the number of bootstrap samples used to construct the ensemble. As the number of bootstraps used increases, so does the stability of the ensemble. The default value of 40 bootstrap rounds is suitable since increasing it to 60 rounds only marginally increases the stability while requiring a 50% increase of the computational effort. On the other hand, the number of bootstraps does not seem to have an effect on the classification performance. Results for the other data sets can be found as supplementary material (figures 3 and 4).

3.2.2 Sensitivity with regard to the number of features to eliminate
The number of features to eliminate does not have an effect on the ensemble methods, neither regarding stability nor classification performance (figures 6 and 7). For the baseline model on the other hand, ranking the features in one step, i.e. removing 100%, yields the best results, both in terms of stability and classification performance. While the effect is only marginal regarding classification performance, the effect has a clear impact on the stability: a single SVM classifier used as a feature ranker largely outperforms the RFE variants that eliminate 20% or 50% of the features. Ensemble methods nevertheless show better performance both in terms of stability and classification, as compared to a single SVM run.

4 DISCUSSION

The present study discusses the robustness of biomarker identification with a concrete focus on microarray experiments on four cancer diagnosis data sets. We stress the importance of the marker stability with respect to sample variation. Such a stability is desirable both for the reproducibility and the easiness of the biological validation of the extracted signature. Our underlying hypothesis is that small changes in the data sampling should not yield dramatic changes in the set of finally selected markers. Stability alone is however not a good quality criterion since it is straightforward to increase stability by always considering some fixed set of markers. The resulting predictive model would however likely be poor at classifying new samples. Our first contribution is an experimental methodology to assess the stability of biomarker lists combined with the predictive performance of classification models built from them. Such an experimental protocol repetitively considers some samples both to select markers and to estimate classifiers from independent samples used to estimate classification performance. In particular it avoids a common optimistic bias of the selection process (Ambroise and McLachlan, 2002). Our protocol also relies on the area under the ROC curve (AUC) and the Kuncheva stability index (KI). AUC is a more convenient metric than classification accuracy to evaluate the predictive performance on data sets with unbalanced class proportions, a common situation for microarray experiments. KI measures to which extent several signatures, typically extracted from different sampling of the data, share features in common while including a correction of such a common selection purely by chance.

Our second contribution is a set of ensemble feature selection methods improving biomarker stability and classification performance. We used RFE as a baseline method because it already offered state-of-the-art classification performance and it is intrinsically multivariate, that is measuring the joined relevance of sets of markers. When decreasing the number of selected features, the stability of RFE tends to degrade while ensemble methods offer significantly better stability. Stability and classification performance are particularly improved for signature sizes as small as 0.5% of the initial feature set. This is particularly convenient since it corresponds to sizes of practical interest (a few tens of genes), for instance, for the design of a diagnosis model.

The present work proposes ensemble methods to improve biomarker stability. Since this is an important issue, it would be interesting to develop additional alternatives to further increase stability. Recent results show that incorporation of prior knownledge in the biomarker selection process (Helleputte and Dupont, 2009b) or model estimation across several related datasets (Helleputte and Dupont, 2009a) are worth investigating along these lines.

ACKNOWLEDGMENTS

Thomas Abeel would like to thank *IWT-Vlaanderen*, Thibault Helleputte thanks the *Fonds pour la formation à la Recherche dans l'Industrie et dans l'Agriculture (FRIA)* and Yvan Saeys thanks the *Research Foundation-Flanders (FWO-Vlaanderen)* for funding their research.

REFERENCES

Alizadeh, A., Eisen, M., Davis, R., Ma, C., Lossos, I., Rosenwald, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., Powell, J., Yang, L., Marti, G., Moore, T., Jr, J. H., Lu, L., Lewis, D., Tibshirani, R., Sherlock, G., Chan, W., Greiner, T., amd J.O. Armitage, D. W., Warnke, R., Levy, R., Wilson, W., Grever, M., Byrd, J., Botstein, D., Brown, P., and Staudt, L. (2000). Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, **403**(3), 503–511.

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., and Levine, A. (1999). Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, **96**(12), 6745–6750.

Ambroise, C. and McLachlan, G. (2002). Selection bias in gene extraction on the basis of microarra gene-expression data. *PNAS*, **99**(10), 6562–6566.

Ben-Hur, A., Ong, C., Sonnenburg, S., Schoelkopf, B., and G. Raetsch, G. (2008). Support vector machines and kernels for computational biology. *PLoS Comput Biol*, **4**(10), e1000173.

Boser, B., Guyon, I., and Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Proc. COLT*, pages 144–152. ACN Press.

Braga-Neto, U. and Dougherty, E. (2004). Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, **20**(3), 374–380.

Dettling, M. (2004). Bagboosting for tumor classification with gene expression data. *Bioinformatics*, **20**(18), 3583–3593.

Dietterich, T. (2000). Ensemble methods in machine learning. In *Proceedings of the 1st International Workshop on Multiple Classifier Systems*, pages 1–15.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, **7**(1), 1–26.

Golub, T., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, **3**, 1157–1182.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, **46**(1-3), 389–422.

Helleputte, T. and Dupont, P. (2009a). Feature selection by transfer learning with linear regularized models. *Lecture Notes in Artificial Intelligence*, **5781**, 533–547.

Helleputte, T. and Dupont, P. (2009b). Partially supervised feature selection with regularized linear models. In *26th International Conference on Machine Learning (ICML)*.

Kalousis, A., Prados, J., and Hilario, M. (2007). Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowl. Inf. Syst.*, **12**(1), 95–116.

Krishnapuram, B., Carin, L., and Hartemink, A. (2004). *Kernel Methods in Computational Biology*, chapter 14: Gene Expression Analysis: Joint Feature Selection and Classifier Design, pages 299–317. MIT Press, Cambridge, MA.

Kuncheva, L. (2007). A stability index for feature selection. In *Proceedings of the 25th International Multi-Conference on Artificial Intelligence and Applications*, pages 309–395.

Long, P. and Berlian Vega, V. (2003). Boosting and microarray data. *Machine Learning*, **52**, 31–44.

Provost, F. and Fawcett, T. (1997). Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *Proc. of the Third Intl. Conf. on Knowledge Discovery and Data Mining*, pages 43–48.

Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**(19), 2507–2517.

Saeys, Y., Abeel, T., and Van de Peer, Y. (2008). Robust feature selection using ensemble feature selection techniques. In *Proceedings of the 25th European Conference on Machine Learning and Knowledge Discovery in Databases, Part II*, pages 313–325.

Schoelkopf, B. and Smola, A. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P., Renshaw, A., D'Amico, A., Richie, J., Lander, E., Loda, M., Kantoff, P., TR, T. G., and Sellers, W. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*, **1**(2), 203–209.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.