BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–9

Microarray-based classification and clinical predictors: On combined classifiers and additional predictive value

Anne-Laure Boulesteix [1,2,*], Christine Porzelius [1,3] and Martin Daumer[1]
*

[1] Sylvia Lawry Centre for MS Research. Hohenlindenerstr. 1. D-81677 Munich (Germany)
[2] Department of Statistics, Ludwig-Maximilians-University of Munich, Ludwigstr. 33. D-80539 Munich (Germany)
[3] Institute of Medical Biometry and Medical Informatics, University Hospital Freiburg. Stefan-Meier-Str. 26. D-79104 Freiburg (Germany)

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: In the context of clinical bioinformatics methods are needed for assessing the additional predictive value of microarray data compared to simple clinical parameters alone. Such methods should also provide an optimal prediction rule making use of all potentialities of both types of data: they should ideally be able to catch subtypes which are not identified by clinical parameters alone. Moreover, they should address the question of the additional predictive value of microarray data in a fair framework.

Results: We propose a novel but simple two-step approach based on random forests and PLS dimension reduction embedding the idea of pre-validation suggested by Tibshirani and colleagues which is based on an internal cross-validation for avoiding overfitting. Our approach is fast, flexible and can be used both for assessing the overall additional significance of the microarray data and for building optimal hybrid classification rules. Its efficiency is demonstrated through simulations and an application to breast cancer and colorectal cancer data.

Availability: Our method is implemented in the freely available R package 'MAclinical' which can be downloaded from http://www.stat.uni-muenchen.de/∼socher/MAclinical.

Contact: boulesteix@slcmsr.org

1 INTRODUCTION

For the last few years, microarray-based outcome prediction, especially classification, has attracted much attention in the statistics, bioinformatics and medical communities. While cancer research is probably the most important field of application of microarray-based prediction, classifiers have also been proposed for other diseases such as multiple sclerosis (Bomprezzi et al., 2003). Classification studies using microarray data only often aim to demonstrate that microarray data are informative to distinguish different types of tissues or patients, e.g., normal from cancer tissues or responders from non-responders. As a by-product of such a study, researchers usually also explore the molecular mechanisms underlying the considered disease by focusing their attention on the most informative genes.

In the context of outcome prediction, some groups of researchers suggest that gene expression data could be used in clinical practice to provide improved diagnosis or prediction (see, e.g. van't Veer et al., 2002). In this case, it is crucial to assess the additional predictive value of gene expression data compared to the available (good) simple clinical predictors. Since they are in general much more difficult and expensive to collect than clinical predictors, gene expression predictors should be used as prediction tools only when they really lead to an accuracy improvement. A problem related to the additional predictive value is outlined by Ntzani and Ioannidis (2003) who state that *'adjustment for other classic predictors of the disease outcome [is] essential'*. This is especially true when the study's aim is to demonstrate the practical benefit of using gene expression predictors in clinical practice, but also in other cases. For instance, suppose that the age and sex distributions are not the same in the two groups that have to be distinguished. If these variables are ignored when performing classification, one may misleadingly conclude that microarray data can separate the two groups very well, whereas the differences in gene expression are in fact due to sex and age differences.

Although taking clinical variables into account may be crucial in the context of microarray-based prediction, this aspect is often either omitted or performed using sub-optimal methods and not adequately described in the medical literature. Hundreds of novel methods have been proposed to deal with the 'small n, large p' problem, but very few statisticians address the question of the additional predictive value of microarray data. We give an overview at the end of this section.

Such clinical parameters may include, e.g. age and sex of the patient, disease duration, relapse rate or tumor grade, depending on the investigated disease. A critical study of breast cancer outcome prediction (Eden et al., 2004) suggests that *' 'good old' clinical markers have similar power in breast cancer prognosis as microarray [...] profilers'* and, more generally, microarray data are suspected to sometimes yield *'noise discovery'* (Ioannidis, 2005). In another context, Hunter et al. (2008) point out that *'letting the genome out of*

*to whom correspondence should be addressed

the bottle' may have perverse effects in the context of genetic tests. Similar results have been obtained in the field of multiple sclerosis and magnetic resonance imaging (MRI). MRI, which has long been considered as an efficient tool for disease course prediction, turns out to show only marginal additional predictive value when it is used in combination with simple clinical parameters including, e.g., relapse history and disease duration (Daumer et al., 2006).

In the present paper, we focus on a standard binary classification problem: the response variable Y to be predicted can take two values $Y = 0$ or $Y = 1$. The term prediction refers to the prediction of the response Y. For example, Y may stand for the development of metastases within a given period of time (yes/no). Note that not all prediction problems can be easily simplified in terms of binary prediction without substantial loss of information and precision. However, class prediction remains the most commonly encountered prediction problem in high-dimensional settings. Our method is easily generalizable to other prediction problems including survival analysis and multicategorical responses.

The answer to the question of the additional predictive value of microarray data is typically binary: 'yes, microarray data improve the classification accuracy yielded by clinical predictors' or 'clinical predictors perform at least as well as gene expression predictors -and are much less expensive'. The second answer may correspond to different situations. Firstly, it is possible that microarray data are not relevant at all for the prediction problem, in which case a usual classifier for high dimensional data gives poor results when applied to microarray data alone. The second scenario is that microarray data are relevant for the prediction problem, but redundant with or weaker than clinical parameters, in which case a usual classifier for high dimensional data yields satisfying results. Note that the term 'redundant' does not imply any causality relationship. Microarray data and clinical data may be redundant because the gene expression influences clinical variables or vice versa, or because both clinical and microarray variables are influenced by common latent unobserved mechanisms. Additional biological knowledge is needed to answer this question, which goes beyond the scope of this article.

In practical studies, the additional predictive value of microarray data is often assessed by using naïve methods. The most simple one is probably subgroup analysis. If one is interested in the predictive value given that a binary predictor is already available, the separate analysis of both subgroups is a natural approach. Considering the small sample sizes in microarray studies and the number of available candidate clinical predictors (typically about 5 to 10), this approach can not be recommended in general. Another simple approach consists of building a classifier based on all predictors, without distinguishing between microarray and clinical variables. This method seems also inappropriate to answer the question of the additional predictive value: even if we have an excellent clinical predictor, it is likely to get lost within the huge amount of microarray variables. Hence, this approach does not treat clinical predictors fairly. The third intuitive approach consists of building two classifiers: one based on clinical parameters, one based on microarray data. The problem is then that the original question of the additional predictive value cannot be answered at all. If both classifiers perform similarly, one does not know whether microarray data do exactly the same as clinical parameters or rather allow to refine the prediction in some way. Hence, the assessment of the additional predictive value of microarray data is not a trivial issue.

A related problem is the construction of complex classifiers combining clinical parameters and high dimensional microarray data. Ideally, such a classifier would

1. show at least as good performance as simpler classifiers using only clinical parameters or only microarray data, respectively,
2. handle different configurations (bad microarray and good clinical predictors, good microarray and bad clinical predictors) by performing correct model selection,
3. neither over-summarize microarray data nor favor them in the final classifier through overfitting mechanisms,
4. handle both categorical and continuous predictors, since many clinical parameters are categorical,
5. decide automatically whether to include microarray data or not, depending on their additional predictive value.

In the literature, some articles address the question of the additional predictive value of microarray data, whereas others propose combined classifiers without answering this question. Here is a brief review.

On the one hand, Tibshirani and Efron (2002) suggest the so-called 'pre-validation' (PV) testing framework whose aim is to determine whether microarray data contribute significantly to the prediction problem, given that clinical parameters are already available. The idea is to summarize microarray data in form of the internally cross-validated predicted probability of class membership, thus avoiding that microarray data are artificially favored. This approach is applied, e.g., in a breast cancer study by Pawitan et al. (2005). Note that the aim of this method is not to construct an optimal classifier combining both types of data.

On the other hand, several authors try to involve clinical parameters in the classifier construction in some way. Dettling and Bühlmann (2004) suggest a statistical approach based on penalized logistic regression handling all types of clinical variables. Gevaert et al. (2006) follow an approach based on Bayesian networks involving two steps (structure step and learning step). A related approach is presented by Sun et al. (2007). It is also based on variable selection, although using a completely different selection procedure. The method by Sun et al. (2007) relies on a wrapper feature selection method called I-RELIEF. They use linear discriminant analysis (LDA) as a class prediction method, which can be an inconvenience in the presence of categorical predictors.

Some of these studies do not appear to use any systematic validation strategy and hence have the pitfalls outlined by Dupuy and Simon (2007), which make their results uninterpretable. Moreover, most of them do not provide any adequate answer to the related question of the additional predictive value of microarray data from a testing point of view because it was not their primary goal. For instance, with methods putting microarray and clinical data together, the latter tend to get lost within the huge amount of microarray variables and are thus not treated fairly from the point of view of the additional predictive value. Methods treating the two groups of variables separately and combining them at the end may also fail partly in the frequent case where clinical and microarray data are highly correlated.

In this article, we present a method which simultaneously i) determines whether microarray data have additional predictive value and

ii) provides a combined classifier fulfilling the five points enumerated above. To the best of our knowledge, there is no other approach treating these two aspects in a common framework. In a very recent article, Binder and Schumacher (2008) address these problems based on a penalized Cox regression approach using componentwise boosting techniques. However, they only address the prediction of survival times. It is still unclear whether such methods would perform well for classification problems in high dimensional settings, which may be more affected by separation and overfitting problems.

According to several independent comparison studies (Man et al., 2004; Boulesteix, 2004; Dai et al., 2006), PLS-based methods range among the best dimension reduction methods for high-dimensional and noisy microarray data in the context of prediction. See Nguyen and Rocke (2002) for the first application of PLS to microarray-based prediction and Boulesteix and Strimmer (2007) for an overview of PLS methods for genomic data.

In this article, we suggest a new approach combining PLS dimension reduction and the principle of pre-validation introduced by Tibshirani and Efron (2002). Random forests (Breiman, 2001) are then applied with both the new components and the clinical variables as predictors. Our proposal contains several novelties: i) the two-step approach involving a dimension reduction step and a classification step for handling the two types of variables, ii) the extension of the pre-validation idea to dimension reduction and prediction, iii) the combination of PLS and random forests which involves several advantages, and iv) a model choice procedure based on the out-of-bag error estimator. The proposed method is described in Section 2 and illustrated in Section 3 through simulations and an application to the breast cancer data by van't Veer et al. (2002) and the colorectal data by Lin et al. (2007).

2 METHODS

Let $\mathbf{X}$ denote the $n \times p$ matrix containing the column-centered expression values of p genes for n patients, while $\mathbf{y}$ denotes the centered vector of classes coded as 0, 1. Similarly, $\mathbf{Z}$ denotes the $n \times q$ matrix giving the values of q clinical parameters for the n patients. In contrast to the gene expression matrix $\mathbf{X}$, $\mathbf{Z}$ may include categorical variables, such as tumor grade or sex of the patient.

In the present section, we first give a short overview of partial least squares (PLS) dimension reduction and random forest classification. In the second subsection, we propose a novel method combining PLS dimension reduction with the idea of pre-validation suggested by Tibshirani and Efron (2002). We then outline the whole procedure consisting of summarizing microarray data in form of pre-validated PLS components and applying random forests to both microarray and clinical variables and address the problem of model choice, especially the choice of the number of PLS components.

2.1 An introduction to PLS and random forests

Partial Least Squares (PLS) methods were developed in connection with path models in the 60s and 70s (Wold, 1966). Statisticians became interested in its application to robust and computationally efficient regression for data with small sample sizes and large number of highly correlated variables some 25 years ago (Martens and Naes, 1989; Stone and Brooks, 1990; Garthwaite, 1994). The following introduction refers to the review by Boulesteix and Strimmer (2007). PLS regression consists of two steps. During the dimension reduction step, the predictors from matrix $\mathbf{X}$ are summarized in form of a small number of linear combinations called 'PLS components'. Subsequently, assuming that the response is continuous, these extracted PLS components are used as predictors in ordinary least squares regression, hence the term 'PLS regression'. When the response is binary, the linear regression step can of course not be carried out. However, it can be shown (Barker and Rayens, 2003) that, if applied to a categorical response, the dimension reduction step is strongly related to principal component analysis performed on the between-group covariance matrix. Hence, it makes sense to perform PLS dimension reduction in this setting.

PLS dimension reduction constructs k mutually orthogonal components as linear combinations of $\mathbf{X}$:

$$\mathbf{T} = \mathbf{XW},$$

where $\mathbf{T}$ is the $n \times k$ matrix of new components $\mathbf{t}_i = (t_{1i}, ..., t_{ni})^T$, for $i = 1, ..., k$, and $\mathbf{W}$ a $p \times k$ matrix of weights satisfying particular optimality criteria. These criteria differ slightly depending on the considered PLS variant. One of the most widely used PLS variant is SIMPLS (de Jong, 1993), in which $\mathbf{W}$ is constructed such that the squared sample covariance between $\mathbf{y}$ and the latent components is maximal under the constraint that the columns $\mathbf{w}_1, ..., \mathbf{w}_k$ of $\mathbf{W}$ are of unit length and the new components $\mathbf{t}_i$ are mutually orthogonal. In mathematical terms the extraction of the subsequent components can be written as

$$\mathbf{w}_i = \arg\max_{\mathbf{w}} Cov^2(\mathbf{Xw}, \mathbf{y}) = \arg\max_{\mathbf{w}} (\mathbf{y}^T \mathbf{Xw})^2 \qquad (1)$$

subject to $\mathbf{w}_i^T \mathbf{w}_i = 1$ and $\mathbf{t}_i^T \mathbf{t}_j = \mathbf{w}_i^T \mathbf{X}^T \mathbf{X} \mathbf{w}_j = 0$. The fast extraction of the weight matrix $\mathbf{W}$ can be carried out using a sequential algorithm given in, e.g., Martens and Naes (1989). By definition, the most informative components are the first ones, but the determination of the best number of components is a difficult task. Some authors (Boulesteix, 2004; Dai et al., 2006) use cross-validation based strategies. In this article, we use the implementation of SIMPLS included in the R package 'plsgenomics' (Boulesteix, 2004; Boulesteix and Strimmer, 2007), function 'pls.regression'.

Although variable selection is not always necessary as a preliminary step to PLS-based classification, some authors argue that it can substantially improve accuracy in the high-dimensional setting (Dai et al., 2006), especially when there are indeed few relevant variables. Many variable selection procedures are available in the literature. One of the most widely used is univariate filtering based on the absolute value of the t-statistic. In the present paper, we stick to this standard approach.

Random forests are introduced by Breiman (2001) and based on the decision tree methodology. In only seven years, they have grown to a major data analysis tool, especially in the context of high-dimensional genetic or genomic data (Strobl et al., 2007). Like bagging (Breiman, 1996), the method is based on the aggregation of classification or regression trees built using bootstrap samples drawn out of the original n observations, in order to make tree-based prediction more robust. In order to make the obtained trees even more different and thus increase their stability and to reduce the computation time, random forests have an additional feature. At each split, a subset of candidate predictors is selected out of the available predictors. The size $mtry$ of the subset, which is a method parameter, is often set to $mtry = \sqrt{p}$, where p is the total number of predictors.

As a model-free approach, the random forest method does not need any distributional assumptions and can be applied to any type of data. In particular, it behaves well with high-dimensional correlated data, see, e.g., Diaz-Uriarte and de Andrés (2006) for an application to microarray-based class prediction. Random forests can also take interactions between variables into account explicitly. Lastly, they are faster than other aggregation methods like bagging, since they do not consider all the available predictors at each split.

Like classification and regression trees, random forests handle all types of responses, in particular multicategorical or censored responses. They also work with all types of predictor variables. However, when predictors do not have the same scale, selection bias may occur using the standard random forest algorithm. It is then recommended to use an alternative version of the random forest method based on conditional inference (Hothorn et al., 2006) implemented in the function 'cforest' of the R package 'party'. Moreover, it can be shown (Strobl et al., 2007) that subsampling (without replacement) is

preferable to the bootstrap when drawing samples out of the n observations at each random forest iteration. In this paper, we follow these recommendations. The only parameters for which we do not use the default settings are i) the number of trees, which is set to 200 instead of 500 for computational reasons, ii) the number of candidate predictors at each split, which we set to $\sqrt{p}$ for consistency with the original R package 'randomForest' implementing the method by Breiman (2001), iii) the threshold defining the stopping criterion (see Hothorn et al. (2006) for more details), which we set to `mincriterion=0` in order to obtain trees with long branches, as commonly recommended for trees used in random forests. In very small data sets (say, $n \leq 30$), one should also modify the parameter `minsplit` controlling the minimal size of nodes to be split. However, our experience shows that this modification is not necessary in data sets of usual size as those considered here. Note that, in contrast to other methods such as penalized logistic regression, the performance of random forests depends only slightly on the choice of parameters and that different settings would yield similar results.

2.2 Pre-validated PLS

Suppose that we construct PLS components as described in Section 2.1, based on a given learning data set. Per definition, these components are likely to be strongly related to the response variable, especially in the case of high dimensional data. Comparing their predictive power to the power of clinical variables in the learning data set would be an unwise strategy: Because of overfitting, there typically will be a bias in favor of the PLS components.

In the present article, we suggest to overcome this problem by extending the pre-validation principle of Tibshirani and Efron (2002) to PLS dimension reduction. Pre-validation is inspired from the well-known cross-validation procedure for evaluation of prediction rules, which consists of partitioning the available sample into distinct subsamples and successively considering each subsample as test data and the remaining subsamples as training data. Unfamiliar readers may refer to the review by Boulesteix et al. (2008) on this subject. Our novel procedure works as follows.

1. Divide the learning data set into G groups. Here, we set $G = 10$, as recommended by Höfling and Tibshirani (2008).
2. Leave one group out and run PLS dimension reduction on the remaining $G - 1$ groups.
3. Compute the PLS components for the left-out group using the derived weight matrix. We denote these PLS components as *pre-validated PLS components*.
4. Repeat steps 2-3 for each of the G groups.

The pre-validated components can then be fairly compared to other variables.

2.3 Summary: Recipe of the analysis

In the present article, we suggest to combine PLS dimension reduction with the random forest methodology in order to take both gene expression and clinical parameters into account when constructing a classifier. Suppose that we have a learning data set L of size n_L (corresponding to $\mathbf{X}_L$, $\mathbf{Z}_L$, $\mathbf{y}_L$) for which we know the response variable. We also have a test data set T of size n_T (corresponding to $\mathbf{X}_T$, $\mathbf{Z}_T$), for which a prediction has to be made.

In clinical practice, the test data set would be a set of patients that have to be predicted. In the context of the validation of research findings, the test data set would be a set of patients for which we also know the response variable, and that are used to assess the prediction accuracy of the combined classifier constructed using the learning data. Note that this scheme is possible only if we have a large enough data set. Otherwise, one may use an evaluation scheme based on, e.g., cross-validation, repeated subsampling, or bootstrap sampling, see Boulesteix et al. (2008) for an overview. In this case, the algorithm is run several times. For example, if leave-one-out cross-validation (LOOCV) is used to assess our combined classifier, one would run the following algorithm for each LOOCV iteration, where the data set $\mathbf{X}_T$ consists of only one observation at each iteration.

The matrix $\mathbf{X}_L$ is assumed to have columns with zero mean and $\mathbf{X}_T$ to be centered by substraction of the columns' means obtained from $\mathbf{X}_L$, as usual in PLS-based prediction (Boulesteix and Strimmer, 2007). Let k denote the maximum allowed number of PLS components, typically $k = 3$ in the binary case. The detailed procedure is as follows.

1. **Cross-validated PLS dimension reduction with learning data set** Construct the $n_L \times k$ matrix of pre-validated PLS components $\tilde{\mathbf{T}}_L$ as follows. For $g = 1, \ldots, G$:

   1a. Carry out variable selection based on $\mathbf{X}_L^{(-g)}$ and $\mathbf{y}_L^{(-g)}$ only, where the superscript '$(-g)$' indicates that the observations from the g-th group have been removed from $\mathbf{X}_L$ and $\mathbf{y}_L$, respectively. This yields an expression matrix $\mathbf{X}_L^{*(-g)}$ with p^* columns, where p^* is the pre-fixed number of selected variables. The g-th group is not taken into account in the variable selection process, because variable selection must be considered as a part of the classifier construction, see, e.g., Dupuy and Simon (2007); Boulesteix (2007).

   1b. Run the PLS dimension reduction procedure with k components on the data matrix $\mathbf{X}_L^{*(-g)}$. This yields the $p^* \times k$ weight matrix $\mathbf{W}_L^{(-g)}$.

   1c. Build the k components for the excluded g-th group as the product $\mathbf{X}_L^{*(g)}\mathbf{W}_L^{(-g)}$, where $\mathbf{X}_L^{*(g)}$ denotes the part of the matrix $\mathbf{X}_L$ corresponding to the g-th group and containing only the p^* variables selected in 1a. Store the product $\mathbf{X}_L^{*(g)}\mathbf{W}_L^{(-g)}$ in the rows of $\tilde{\mathbf{T}}_L$ corresponding to the g-th group.

2. **Classifier construction** Construct a random forest using the columns of the matrices $\tilde{\mathbf{T}}_L$ and $\mathbf{Z}_L$ as predictors and $\mathbf{y}_L$ as response.
3. **PLS dimension reduction with the test data set** Carry out variable selection based on $\mathbf{X}_L$ and $\mathbf{y}_L$ only, yielding again p^* selected variables. For the reduced test data matrix $\mathbf{X}_T^*$ consisting of the p^* selected variables, compute the matrix $\mathbf{T}_T$ of PLS components as follows.

   3a. Run the PLS dimension reduction procedure with k components on the whole learning data matrix $\mathbf{X}_L^*$ of size $n_L \times p^*$, yielding the $p^* \times k$ weight matrix $\mathbf{W}_L$.

   3b. Build the PLS components for the test data set as $\mathbf{T}_T = \mathbf{X}_T^*\mathbf{W}_L$.

4. **Prediction** Apply the random forest constructed in step 2 to the prediction of the test observations using the matrices $\mathbf{T}_T$ and $\mathbf{Z}_T$. For each test observation one obtains a prediction $\hat{Y}$ for the class membership.

This procedure is summarized as a flow chart in Figure 1.

2.4 Model choice and additive predictive value of microarray data

In this section, we show how the out-of-bag (OOB) error estimator (Breiman, 2001) yielded as a by-product when growing a random forest can be used both for the choice of the number of PLS components and for answering the question of the additive predictive value of microarray data.

The OOB error estimator works as follows. When growing each tree of the random forest, 36.8% of the n observations are put aside and not used for choosing the splits (this default setting of the function `cforest`, which stems from bootstrap sampling). These observations are called *out-of-bag* observations. After all the trees are constructed, a pseudo-prediction can be made for each of the n observations using only the trees that did not use it for training, i.e. the trees for which it was an out-of-bag observation. The OOB error rate is then computed by comparison of the n pseudo-predictions with the true classes. Note that, in contrast to in-bag error estimation, this

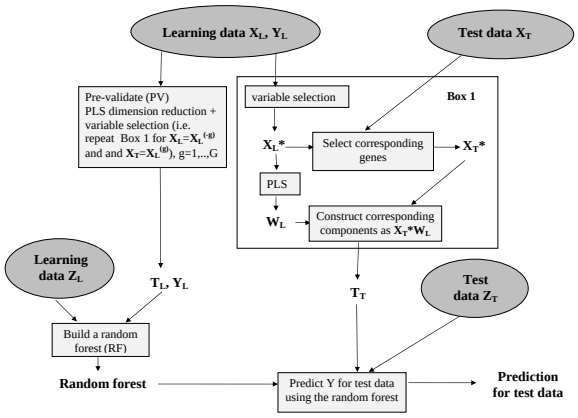

Fig. 1. Schematic representation of the classification methods based on pre-validated dimension reduction and random forests, using both microarray predictors ($\mathbf{X}_L$ for the learning set and $\mathbf{X}_T$ for the test set) and clinical predictors ($\mathbf{Z}_L$ for the learning set and $\mathbf{Z}_T$ for the test set). Predictor data sets are represented in ellipses, actions are represented in boxes, output as simple text.

procedure overcomes overfitting problems, since the predicted observations were not used for training the corresponding predicting trees. The OOB error estimator can be used for comparing the prediction accuracy of several random forests. An interesting application in the present context is the choice of the number of components, and, if zero is considered as a possible candidate for the number of components, the question of the additive predictive value of microarray data. To do this, we suggest to replace step 2 of the procedure outlined above by a modified version 2* as follows.

2*. **Classifier construction**

- For $l = 0, \ldots, k$, construct a random forest using the l first columns of the matrix $\tilde{\mathbf{T}}_L$ and the matrix $\mathbf{Z}_L$ as predictors and $\mathbf{y}_L$ as response.
- Compute the OOB error for each of the $k + 1$ constructed forests.
- Select the number of components k^* yielding the forest with the smallest OOB error.

The number k is then replaced by k^* in the following step of the procedure (Step 3: 'PLS dimension reduction with the test data set'). Note that this procedure is much faster than cross-validation for the choice of the number of components, since the OOB estimator is a by-product of the random forest algorithm. In the rest of this article, this method is denoted as PLS+RF.

If $k^* = 0$, we conclude that microarray data do not have any predictive value compared to clinical variables alone. If $k^* > 0$, it is possible to roughly evaluate the significance of microarray data by computing confidence intervals for the calculated OOB errors, where the sample size is given as the size of the training set. This procedure does not yield a rigorous statistical test, since independence of the observations is not warranted. However, the resulting confidence intervals should give the order of magnitude of the corresponding differences in accuracy.

The whole procedure is implemented in the R package 'MAclinical'. The current version that was used for this paper is available from http://www.stat.uni-muenchen.de/∼socher/MAclinical. We plan to send a refined version of this package to the Comprehensive R Archive Network.

3 RESULTS AND DISCUSSION

The analyzes described below can be reproduced using the scripts available from http://www.stat.uni-muenchen.de/∼socher/MAclinical.

3.1 Simulations

The aim of this simulation study is to compare the performance of our approach to related approaches based on clinical and/or microarray variables. Several data structures are considered: different predictive powers for the microarray variables, different powers for the clinical variables, and different class structures. By different class structures, we mean that we examine two settings: i) a 'redundant' setting where the microarray and clinical variables are generated using exactly the same model, thus discriminating the classes in the same way and giving 'redundant' information, and ii) a 'non-redundant' setting, where observations from class $Y = 1$ are assumed to form two distinct subgroups: one of the subgroups can be discriminated from the other one and from $Y = 0$ by microarray data, whereas the second one is discriminated by clinical variables. The corresponding data generating processes are detailed below.

In the first setting (redundant setting), the random variables Y, $X_1, \ldots, X_p$ and $Z_1, \ldots, Z_q$ have the following joint distribution. The binary response Y follows a binomial distribution with $P(Y = 1) = 0.5$. A total of $p^* < p$ microarray variables are relevant for class prediction. Each microarray variable X_j ($j = 1, \ldots, p$) is generated as

$$X_j = \mu_{X_j} \cdot Y + e_j, \tag{2}$$

and each clinical variable Z_s ($s = 1, \ldots, q$) as

$$Z_s = \mu_{Z_s} \cdot Y + f_s, \tag{3}$$

where μ_{X_j} ($j = 1, \ldots, p$) and μ_{Z_s} ($s = 1, \ldots, q$) are constant parameters controlling the amount of predicting power of the microarray and clinical variables, respectively, and the terms e_j ($j = 1, \ldots, p$) and f_s ($s = 1, \ldots, q$) are independent random errors following a standard normal distribution.

In the present simulation, we set μ_{Z_s} to the same value $\mu_{Z_s} = \mu_Z$ for all clinical variables and consider different values of μ_Z successively: $\mu_Z = 0$ (no power), $\mu_Z = 1$ (moderate power) and $\mu_Z = 3$ (strong power). Similarly, μ_{X_j} is set to the constant μ_X for the p^* genes $X_1, \ldots, X_{p^*}$ (with $p^* < p$) and to zero for the remaining genes $X_{p^*+1}, \ldots, X_p$. Similarly to μ_Z, the parameter μ_X takes different values successively: $\mu_X = 0, 0.5, 1$. In the present study, the total number of genes is set to $p = 1000$ and the number p^* of relevant genes to $p^* = 50$. We denote this simulation setting as redundant, because the discrimination mechanism is the same for microarray and clinical variables.

For all parameter combinations (μ_X, μ_Z), we draw n_L and n_T i.i.d. observations forming the learning set and test set, respectively. Here, n_L is set to $n_L = 50$, n_T is set to $n_T = 450$ in order to obtain accurate estimates of the error rate. A total of $N_{iter} = 100$ data sets are simulated for each parameter setting. The optimal number of PLS components is selected from $k = 0, 1, 2, 3$.

We compare the pre-validated PLS+RF method based on both microarray and clinical variables ('pls-pv+rf/xz', with 10-fold PV) to simpler related approaches, in order to determine the effect of pre-validation and to answer the question whether the combined classifiers perform as well as classifiers based on microarray data only or clinical variables only. The considered approaches are i)

PLS+RF based on both microarray and clinical variables without pre-validation ('pls+rf/xz'), ii) pre-validated PLS+RF based on microarray data only with 10-fold PV ('pls-pv+rf/x'), iii) PLS+RF based on microarray data only without pre-validation ('pls+rf/x'), and iv) RF based on clinical variables only ('rf/z'). As an additional comparison, we also apply standard approaches used when dealing with only one type of predictors: logistic regression for clinical predictors ('log/z') and the Support Vector Machines (SVM) method for microarray data ('svm/x'), which is well-established as one of the most accurate procedures in this setting (Statnikov et al., 2005). We use the R package 'e1071', with linear kernel and cost set to the default value 1. In order to make clear that we do not 'tune' our method artificially (which would yield an unfair comparison), let us mention that we additionally applied the two pre-validation approaches ('pls-pv+rf/xz' and 'pls-pv+rf/x') with leave-one-out pre-validation instead of 10-fold pre-validation (data not shown). However, the results were not different from those obtained with 10-fold pre-validation. Hence, we stick to 10-fold, following Höfling and Tibshirani (2008).

For each iteration, a classifier is built based on the learning data set only, with the seven methods outlined above successively. The classifiers are then evaluated based on the corresponding test set and the error rate is estimated as the mean proportion of misclassified observations. In this simulation, we do not perform any preliminary variable selection, as suggested by Boulesteix (2004, 2006) in the case of relatively large signal to noise ratios. In real data analysis, one could of course try to improve classification accuracy by preliminary variable selection. This step was omitted in the simulation for computational reasons. Similarly, correlations between genes and/or clinical variables do not seem to affect the results noticeably (data not shown).

As can be seen from Table 1, pre-validation improves classification accuracy noticeably, especially when clinical parameters are good predictors (i.e., for $\mu_Z = 1$ or $\mu_Z = 3$). Since PLS components without pre-validation usually overfit the training data, they are artificially preferred to clinical parameters in the split selection procedure. Moreover, trees are then likely to have longer irrelevant branches. The performance of the 'pls-pv+rf/xz' approach is slightly lower than the performance of 'pls-pv+rf/x' in the case of non-predictive clinical variables, but as good as the 'rf/z' approach in all cases, even when microarray data are not predictive. The comparison to the standard logistic regression and to SVMs reveals interesting features. In Table 1, logistic regression with clinical parameters performs better than the 'pls-pv+rf/xz' approach only when clinical parameters are more predictive than microarray data. Except for the case $\mu_X = 0, \mu_Z = 1$ (0.165 ± 0.03 vs 0.216 ± 0.03), this difference is minimal. Note that in this case, random forests with clinical variables only do not perform better than 'pls-pv+rf/xz'. The difference between random forests and logistic regression can be explained by the linear structure of our simulated data: for this simulation setting the flexibility of random forests is not an advantage, in contrast to the non-redundant setting sketched below. SVMs perform approximately as well as 'pls-pv+rf' in the case of non-informative clinical variables, but worse in all other cases.

As an illustration of the model selection scheme proposed in Section 2.4 and its ability to assess the additional predictive value of microarray data, we also compute i) the mean OOB error over the 100 subsampling runs obtained with $k^* = 0$ and $k^* = 1$ PLS component, and ii) the percentage of runs for which at least one PLS component is selected (i.e. $k^* > 0$), for the novel 'pls-pv+rf/xz' method. As can be seen from Table 2, the proportion of simulation runs with $k^* > 0$ selected PLS components increases drastically with μ_X, but this increase also depends on μ_Z. For a fixed μ_X, the proportion of runs with $k^* > 0$ is much lower with informative clinical variables ($\mu_Z = 1, 3$) than with non-informative clinical variables. This can be explained as follows: if clinical variables perform well, it is more difficult for microarray data to yield accuracy improvement.

The simulation design outlined above corresponds to the case where both microarray and clinical variables discriminate the two response classes in the same way, for instance because both of them

Method	μ_Z	$\mu_X = 0$	$\mu_X = 0.5$	$\mu_X = 1$
pls-pv+rf/xz	0	0.50 ± 0.02	0.33 ± 0.07	0.04 ± 0.03
pls+rf/xz	0	0.50 ± 0.01	0.48 ± 0.04	0.44 ± 0.07
pls-pv+rf/x	0	0.50 ± 0.02	$0.29 \pm 0.06^*$	$0.03 \pm 0.02^*$
pls+rf/x	0	0.50 ± 0.01	0.48 ± 0.04	0.44 ± 0.07
svm/x	0	0.50 ± 0.02	0.30 ± 0.05	0.05 ± 0.03
rf/z	0	0.50 ± 0.02	–	–
log/z	0	0.50 ± 0.02	–	–
pls-pv+rf/xz	1	0.22 ± 0.04	$0.19 \pm 0.04^*$	$0.03 \pm 0.02^*$
pls+rf/xz	1	0.43 ± 0.08	0.42 ± 0.09	0.39 ± 0.10
rf/z	1	0.22 ± 0.04	–	–
log/z	1	$0.17 \pm 0.03^*$	–	–
pls-pv+rf/xz	3	0.01 ± 0.01	$0.01 \pm 0.01^*$	$0.01 \pm 0.01^*$
pls+rf/xz	3	0.05 ± 0.05	0.05 ± 0.05	0.05 ± 0.04
rf/z	3	0.01 ± 0.02	–	–
log/z	3	$0.003 \pm 0.00^*$	–	–

Table 1. Redundant setting. Mean error rate and standard deviation (over 100 simulation runs) for the seven class prediction methods: 'pls-pv+rf/xz', 'pls+rf/xz', 'pls-pv+rf/x', 'pls+rf/x', 'svm/x', 'rf/z', 'log/z' with different powers for the microarray variables ($\mu_X = 0, 0.5, 1$) and clinical variables ($\mu_Z = 0, 1, 3$). The symbol * indicates the best performance for each setting. Gray figures correspond to random predictors not correlated to the class response Y. Summary: Our method is at least as good as the other approaches in almost all settings.

		$\mu_X = 0$	$\mu_X = 0.5$	$\mu_X = 1$
$\mu_Z = 0$	OOB $k = 0$	0.50	0.50	0.50
	OOB $k = 1$	0.50	0.32	0.04
	% $k^* > 0$	65	96	100
$\mu_Z = 1$	OOB $k = 0$	0.23	0.23	0.23
	OOB $k = 1$	0.23	0.19	0.04
	% $k^* > 0$	67	83	100
$\mu_Z = 3$	OOB $k = 0$	0.01	0.01	0.01
	OOB $k = 1$	0.01	0.01	0.01
	% $k^* > 0$	36	40	48

Table 2. Novel PLS-PV+RF method. Mean **OOB error** over the 100 simulations runs with $k = 0$ and $k = 1$ PLS component and percentage of simulation runs yielding $k^* > 0$ (i.e. where prediction accuracy with microarray data was better than without microarray data).

Method	μ_Z	$\mu_X = 0$	$\mu_X = 0.5$	$\mu_X = 1$
pls-pv+rf/xz	0	0.50 ± 0.02	$0.48 \pm 0.03^*$	0.39 ± 0.08
pls+rf/xz	0	0.50 ± 0.01	0.49 ± 0.01	0.49 ± 0.02
pls-pv+rf/x	0	0.50 ± 0.02	0.48 ± 0.04	$0.35 \pm 0.06^*$
pls+rf/x	0	0.50 ± 0.01	0.49 ± 0.01	0.49 ± 0.02
svm/x	0	0.50 ± 0.02	0.46 ± 0.03	0.37 ± 0.04
rf/z	0	0.50 ± 0.02	–	–
log/z	0	0.50 ± 0.02	–	–
pls-pv+rf/xz	1	0.39 ± 0.04	$0.38 \pm 0.04^*$	$0.30 \pm 0.07^*$
pls+rf/xz	1	0.49 ± 0.03	0.49 ± 0.03	0.48 ± 0.03
rf/z	1	0.39 ± 0.04	–	–
log/z	1	$0.37 \pm 0.04^*$	–	–
pls-pv+rf/xz	3	$0.29 \pm 0.04^*$	$0.29 \pm 0.04^*$	$0.19 \pm 0.07^*$
pls+rf/xz	3	0.48 ± 0.04	0.48 ± 0.04	0.47 ± 0.05
rf/z	3	0.30 ± 0.04	–	–
log/z	3	0.32 ± 0.04	–	–

Table 3. Non-redundant setting. Mean error rate and standard deviation (over 100 simulation runs) for the seven class prediction methods: 'pls-pv+rf/xz', 'pls+rf/xz', 'pls-pv+rf/x', 'pls+rf/x', 'svm/x', 'rf/z', 'log/z' with different powers for the microarray variables ($\mu_X = 0, 0.5, 1$) and clinical variables ($\mu_Z = 0, 1, 3$). The symbol * indicates the best performance for each setting. Gray figures correspond to random predictors not correlated to the class response Y. Summary: Our method is at least as good as the other approaches in all settings.

are influenced by the same underlying mechanism. They give essentially redundant information. In practice, investigators often hope that microarray data give additional (i.e. non-redundant) information, for instance, by correctly predicting a particular group that is difficult to predict with clinical predictors only. In the rest of this section, a variant of the above simulation design is applied to investigate the behavior of the different methods in an ideal extreme case where clinical and microarray predictors are perfectly complementary. The observations with $Y = 1$ are assumed to come from two underlying classes $1a$ and $1b$. The microarray variables are drawn to separate $1a$ from the rest, whereas the clinical variables separate $1b$ from the rest. The underlying model is the same as in Eq. (2) and (3), except that Y is replaced by the binary variables $Y^{(a)}$ and $Y^{(b)}$ defined as $Y^{(a)} = 1$ if $Y = 1a$ and 0 otherwise, and $Y^{(b)} = 1$ if $Y = 1b$ and 0 otherwise, respectively.

In this non-redundant setting, the two-step 'pls-pv+rf/xz' approach performs almost uniformly better than all other methods (see Table 3). This is not surprising, since the simulation setting can be seen as an extreme case where the combination of two types of predictors using tree-based methodologies is expected to work well. However, this case is very important in the context of the additional predictive value. Indeed, by additional predictive value, one often implicitly means that microarray data can predict disease subtypes which are wrongly classified by clinical parameters. Note that the overall performance of all methods decreases dramatically compared to the redundant setting. This is because each observation is discriminated by both clinical and microarray variables in the redundant setting, but by only one of the two types of variables (either clinical or microarray) in the non-redundant setting.

3.2 Application to breast cancer data

This widely-used benchmark data set gives the expression levels of 22483 genes for 78 breast cancer patients, of which 34 have poor prognosis and 44 have good prognosis (van't Veer et al., 2002). It can be downloaded from the article webpage. The data set prepared as described in the original manuscript (only genes that show 2-fold differential expression and p-value for a gene being expressed < 0.01 in more than 5 samples are retained, yielding 4348 genes) is included in the R package 'DENMARKLAB' (Fridlyand and Yang, 2004), which we use in the article. The available clinical variables are age (metric), tumor grade (ordinal), estrogen receptor status (binary), progesterone receptor status (binary), tumor size (metric) and angioinvasion (binary).

The classification accuracy is evaluated using the common repeated subsampling method which is a variant of cross-validation also denoted as *Monte-Carlo-cross-validation*, see Molinaro et al. (2005); Boulesteix et al. (2008) for more details. In a nutshell, instead of splitting the original data set consisting of 78 observations into, say, 5, 10 or n subsets (like in standard cross-validation), we repeatedly split it into a learning set and a test set according to the ratio 4:1. Variable selection is carried out based on the absolute value of the t-statistic, using the learning set only as commonly recommended (Boulesteix, 2007; Dupuy and Simon, 2007). The method described in Section 2.3 is then applied to the learning and test sets including the step for the optimization of the number of components as described in Section 2.4. The error rates are estimated as the proportion of misclassified test observations. The whole procedure is repeated 100 times and the error rates are averaged. This approach usually leads to more stable results than standard cross-validation, because it is based on a larger number of iterations. Table 4 gives the obtained mean error rates with the seven methods described in Section 3.1, for different numbers of variables in the range of the number of genes used in the original signature proposed by van't Veer et al. (2002).

It can be seen from Table 4 that microarray data do not noticeably improve the prediction accuracy yielded by clinical parameters alone, which corroborates the findings of Eden et al. (2004). This result is confirmed by considering the mean OOB error rates obtained with the different numbers of components. For each of the 100 iterations, we estimate the 95%-confidence interval for the difference of the OOB misclassification rates obtained with $k = 0$ and $k = 1$, respectively. The sample size is 62 for both rates, since each learning set contains $0.8 \times 78 \approx 62$ observations. For $p^* = 100$, the lower bound of the obtained confidence interval exceeds zero for only 9% of the 100 iterations, which suggests that the microarray data do not contribute significantly to the prediction. Similar conclusions are obtained with $p^* = 20, 100, 200$.

3.3 Application to colorectal cancer data

This Affymetrix data set described by Lin et al. (2007) gives the expression levels of 16041 genes for 29 good outcome patients and 26 poor outcome patients with colorectal cancer. In addition to microarray data, the two variables sex and age are available. The data are prepared as described in Lin et al. (2007). Gene expression data are expected to have better predictive power than the variables sex and age which typically yield relatively poor prediction accuracy in the case of cancer.

Method	$p^* = 20$	$p^* = 50$	$p^* = 100$	$p^* = 200$
pls-pv+rf/xz	0.30 ± 0.11	0.31 ± 0.11	0.30 ± 0.12	0.31 ± 0.11
pls+rf/xz	0.30 ± 0.11	0.30 ± 0.11	0.31 ± 0.10	0.35 ± 0.12
pls-pv+rf/x	0.41 ± 0.12	0.42 ± 0.12	0.43 ± 0.11	0.43 ± 0.12
pls+rf/x	0.35 ± 0.11	0.36 ± 0.11	0.37 ± 0.10	0.39 ± 0.11
svm/x	0.40 ± 0.10	0.40 ± 0.10	0.40 ± 0.10	0.40 ± 0.10
rf/z	0.29 ± 0.11	–	–	–
log/z	0.30 ± 0.10	–	–	–

Table 4. Mean classification error rate and standard deviation for the **breast cancer** data. The error rate is estimated by 100 subsampling iterations with splitting ratio 4:1.

p^*		20	50	100	200
Breast	$k = 0$	0.30	0.30	0.30	0.30
	$k = 1$	0.30	0.30	0.30	0.30
	$k = 2$	0.29	0.29	0.30	0.30
	$k = 3$	0.30	0.30	0.30	0.30
	$\%k > 0$	67	70	64	64
Colorectal	$k = 0$	0.56	0.56	0.56	0.56
	$k = 1$	0.43	0.40	0.38	0.38
	$k = 2$	0.39	0.37	0.34	0.35
	$k = 3$	0.41	0.39	0.35	0.36
	$\%k > 0$	90	94	95	95

Table 6. Novel PLS-PV+RF method. **Mean OOB error** over the 100 subsampling runs with $k = 0, 1, 2, 3$ PLS component and **percentage of simulation runs yielding** $k^* > 0$ (i.e. where prediction accuracy with microarray data was better than without microarray data).

Method	$p^* = 20$	$p^* = 50$	$p^* = 100$	$p^* = 200$
pls-pv+rf/xz	0.41 ± 0.15	0.40 ± 0.16	0.35 ± 0.15	0.36 ± 0.16
pls+rf/xz	0.33 ± 0.13	0.32 ± 0.13	0.31 ± 0.12	0.30 ± 0.14
pls-pv+rf/x	0.37 ± 0.14	0.36 ± 0.15	0.33 ± 0.13	0.33 ± 0.13
pls+rf/x	0.32 ± 0.13	0.33 ± 0.12	0.32 ± 0.13	0.31 ± 0.14
svm/x	0.37 ± 0.13	0.37 ± 0.13	0.37 ± 0.13	0.37 ± 0.13
rf/z	0.56 ± 0.15	–	–	–
log/z	0.57 ± 0.13	–	–	–

Table 5. Mean classification error rate and standard deviation for the **colorectal cancer** data. The error rate is estimated by 100 subsampling iterations with splitting ratio 4:1.

The analysis design is the same as for the van't Veer data. As can be seen from the results given in Table 5, the microarray data now have predictive power whereas the two variables age and sex do not. Unsurprisingly, involving the uninformative variables age and sex (methods 'pls-pv+rf/xz' and 'pls+rf/xz') slightly decreases the prediction accuracy compared to the methods without clinical variables 'pls-pv+rf/x' and 'pls+rf/x', but the performance of the approach 'pls-pv+rf/xz' remains comparable to the performance of the standard good performing SVM method. Hence, our approach shows overall good performance in very different situations.

As an illustration of the model selection scheme embedded in the PLS-PV+RF method, we show i) the mean OOB error over the 100 subsampling runs obtained with each number of PLS components, and ii) the percentage of runs for which at least one PLS component is selected ($k^* > 0$), as in the simulation study. As can be seen from Table 6, our method correctly selects at least one PLS component in most runs ($\geq 90\%$) for the colorectal data, i.e. much more often than for the van't Veer data. This result is in agreement with the mean OOB obtained with each number of components. Whereas the mean OOB does not depend on the number of PLS components for the van't Veer data ($OOB \approx 0.30$ for all p^* and all k), it decreases substantially between $k = 0$ and $k = 1$ for the colorectal data, with a further slight decrease from $k = 1$ to $k = 2$.

4 CONCLUSION

We have presented a simple two-step approach based on well-established data analysis tools (PLS and random forests) combined with the pre-validation principle by Tibshirani and Efron (2002). This procedure can simultaneously determine whether microarray data have additional predictive value and provide a combined classifier fulfilling the six points enumerated above. Our fast, simple and flexible method is implemented in the R package 'MAclinical'. Its ability to yield efficient hybrid prediction rules in various settings (good/bad microarray/clinical predictors) was demonstrated in both simulations and real data analysis. In particular, our approach does not seem to overestimate the predictive value of microarray data and yields good prediction accuracies when microarray and clinical parameters do not give redundant predictive information. Let us further mention that as far as computation time is concerned our novel method is similar to SVM, i.e. very fast. In their study of pre-validation, Höfling and Tibshirani (2008) point out substantial biases occurring when pre-validated predicted probabilities are tested for significance in linear regression. In our study, the OOB error slightly decreased with the number of included pre-validated PLS components in the non-informative case, but this trend was minimal. This potential slight bias in model selection, which is probably due to the mechanism outlined by Höfling and Tibshirani (2008) (roughly speaking, observations are not independent anymore in cross-validation), should be addressed in future research.

Note that the OOB error used in this article for the choice of the number of PLS components can also be used for the choice of the number of genes, similarly to the procedure for the choice of the number of PLS components. This problem is ignored by many authors, who compare the different number of genes post-hoc, i.e. after completing the evaluation procedure. Tuning the number of genes may lead to optimistic bias, not only in the context considered in this article.

The extension of our method to other prediction problems (regression, survival analysis, multicategorical classification) is straightforward. Unlike other common approaches such as logistic regression, it i) does not require any distributional assumption or a specific type

of relationship (e.g., linearity) between the response and the predictors, ii) does not need any limitation of the number of clinical variables, which is useful in the case of small samples, iii) can cope with separated classes (in this respect, the aggregation of trees built on perturbed data sets is an advantage), and iv) can handle interactions, for instance interactions between microarray and clinical predictors, thus potentially identifying subtypes that are not caught by clinical data alone.

FUNDING AND ACKNOWLEDGMENTS

ALB was sponsored by the Porticus Foundation in the context of the International School for Technical Medicine and Clinical Bioinformatics. We thank J. Friederichs and B. Holzmann for making the colorectal cancer data set available. We are also grateful to Casimir Kulikowski and the three referees for their constructive comments.

REFERENCES

Barker, M., Rayens, W., 2003. Partial least squares for discrimination. Journal of Chemometrics 17, 166–173.

Binder, H., Schumacher, M., 2008. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. BMC Bioinformatics 9, 14.

Bomprezzi, R., Ringnér, M., Kim, S., et al., 2003. Gene expression profile in multiple sclerosis patients and healthy controls: identifying pathways relevant to disease. Human Molecular Genetics 12, 2191–2199.

Boulesteix, A.-L., 2004. PLS dimension reduction for classification with microarray data. Statistical Applications in Genetics and Molecular Biology 3, 33.

Boulesteix, A.-L., 2006. Reader's reaction to 'Dimension reduction for classification with gene expression microarray data' by Dai et al. (2006)'. Statistical Applications in Genetics and Molecular Biology 5, 16.

Boulesteix, A.-L., 2007. WilcoxCV: An efficient R package for variable selection in cross-validation. Bioinformatics 23, 1702–1704.

Boulesteix, A.-L., Strimmer, K., 2007. Partial least squares: A versatile tool for the analysis of high-dimensional genomic data. Briefings in Bioinformatics 8, 32–44.

Boulesteix, A.-L., Strobl, C., Augustin, T., Daumer, M., 2008. Evaluating microarray-based classifiers: An overview. Cancer Informatics 4, 77–97.

Breiman, L., 1996. Bagging predictors. Machine Learning 24, 123–140.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Dai, J. J., Lieu, L., Rocke, D., 2006. Dimension reduction for classification with gene expression data. Statistical Applications in Genetics and Molecular Biology 5, 6.

Daumer, M., Hapfelmeier, A., Neuhaus, A., Ebers, G., 2006. The additional predictive value of magnetic resonance imaging for the prediction of future relapses if relapse history is available. Multiple Sclerosis 12, S46–S47.

de Jong, S., 1993. SIMPLS: An alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems 18, 251–253.

Dettling, M., Bühlmann, P., 2004. Finding predictive gene groups from microarray data. Journal of Multivariate Analysis 90, 106–131.

Diaz-Uriarte, R., de Andrés, S. A., 2006. Gene selection and classification of microarray data using random forests. BMC Bioinformatics 7, 3.

Dupuy, A., Simon, R., 2007. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. Journal of the National Cancer Institute 99, 147–157.

Eden, P., Ritz, C., Rose, C., Fernö, M., Peterson, C., 2004. "Good old" clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers. European Journal of Cancer 40, 1837–1841.

Fridlyand, J., Yang, J. Y. H., 2004. DENMARKLAB R package. Advanced microarray data analysis: Class discovery and class prediction, http://genome.cbs.dtu.dk/courses/norfa2004/Extras/DENMARKLAB.zip.

Garthwaite, P. H., 1994. An interpretation of partial least squares. Journal of the American Statistical Association 89, 122–127.

Gevaert, O., de Smet, F., Timmermann, D., Moreau, Y., de Moor, B., 2006. Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. Bioinformatics 22, e184–e190.

Höfling, H., Tibshirani, R., 2008. A study of pre-validation. Annals of Applied Statistics (in press).

Hothorn, T., Hornik, K., Zeileis, A., 2006. Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics 15 (3), 651–674.

Hunter, D. J., Khoury, M. J., Drazen, J. M., 2008. Letting the genome out of the bottle – Will we get our wish? New England Journal of Medicine 358, 105–107.

Ioannidis, J. P., 2005. Microarrays and molecular research: noise discovery? The Lancet 365, 488–492.

Lin, Y. H., J., J. F., Black, M. A., Mages, J., Rosenberg, R., Guilford, P. J., Phillips, V., Thompson-Fawcett, M., Kasabov, N., Toro, T., Merrie, A. E., van Rij, A., Yoon, H. S., McCall, J. L., Siewert, J. R., Holzmann, B., Reeve, A. E., 2007. Multiple gene expression classifiers from different array platforms predict poor prognosis of colorectal cancer. Clinical Cancer Research 13, 498–507.

Man, M. Z., Dyson, G., Johnson, K., Liao, B., 2004. Evaluating methods for classifying expression data. Journal of Biopharmaceutical Statistics 14, 1065–1084.

Martens, H., Naes, T., 1989. Multivariate Calibration. Wiley, New York.

Molinaro, A., Simon, R., Pfeiffer, R. M., 2005. Prediction error estimation: a comparison of resampling methods. Bioinformatics 21, 3301–3307.

Nguyen, D. V., Rocke, D., 2002. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics 18, 39–50.

Ntzani, E. E., Ioannidis, J. P. A., 2003. Predictive ability of DNA microarrays for cancer outcomes and correlates: An empirical assessment. The Lancet 362, 1439–1444.

Pawitan, Y., Bjöhle, J., Amler, L., Borg, A.-L., Egyhazi, S., Hall, P., Han, X., Holmberg, L., Huang, F., Klaar, S., Liu, E. T., Miller, L., Nordgren, H., Ploner, A., Sandelin, K., Shaw, P. M., Smeds, J., Skoog, L., Wedren, S., Bergh, J., 2005. Gene expression profiling spares early breast cancer patients from adjuvant therapy: Derived and validated in two population-based cohorts. Breast Cancer Research 7, R953–R964.

Statnikov, A., Aliferis, C. F., Tsamardinos, I., Hardin, D., Levy, S., 2005. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 21, 631–643.

Stone, M., Brooks, R. J., 1990. Continuum regression: Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal component regression. J. Roy. Stat. Soc. B 52, 237–269.

Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8, 25.

Sun, Y., Goodison, S., Li, J., Liu, L., Farmerie, W., 2007. Improved breast cancer prognosis through the combination of clinical and genetic markers. Bioinformatics 23, 30–37.

Tibshirani, R., Efron, B., 2002. Pre-validation and inference in microarrays. Statistical Applications in Genetics and Molecular Biology 1, 1.

van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., Friend, S. H., 2002. Gene expression profiling predicts clinical outcome of breast cancer. Nature 415, 530–536.

Wold, H., 1966. Estimation of principal components and related models by iterative least squares, in: P. R. Krishnaiah (Ed.), Multivariate Analysis. Academic Press, New York.