BIOINFORMATICS

Vol. 20 no. 17 2004, pages 3185–3195
doi:10.1093/bioinformatics/bth383

Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction

Nathalie Pochet*, Frank De Smet, Johan A. K. Suykens and Bart L. R. De Moor

ESAT-SCD (SISTA), K.U. Leuven, Kasteelpark Arenberg 10, 3001 Leuven-Heverlee, Belgium

Received on February 17, 2004; revised on April 29, 2004; accepted on June 23, 2004
Advance Access publication July 1, 2004

ABSTRACT

Motivation: Microarrays are capable of determining the expression levels of thousands of genes simultaneously. In combination with classification methods, this technology can be useful to support clinical management decisions for individual patients, e.g. in oncology. The aim of this paper is to systematically benchmark the role of non-linear versus linear techniques and dimensionality reduction methods.

Results: A systematic benchmarking study is performed by comparing linear versions of standard classification and dimensionality reduction techniques with their non-linear versions based on non-linear kernel functions with a radial basis function (RBF) kernel. A total of 9 binary cancer classification problems, derived from 7 publicly available microarray datasets, and 20 randomizations of each problem are examined.

Conclusions: Three main conclusions can be formulated based on the performances on independent test sets. (1) When performing classification with least squares support vector machines (LS-SVMs) (without dimensionality reduction), RBF kernels can be used without risking too much overfitting. The results obtained with well-tuned RBF kernels are never worse and sometimes even statistically significantly better compared to results obtained with a linear kernel in terms of test set receiver operating characteristic and test set accuracy performances. (2) Even for classification with linear classifiers like LS-SVM with linear kernel, using regularization is very important. (3) When performing kernel principal component analysis (kernel PCA) before classification, using an RBF kernel for kernel PCA tends to result in overfitting, especially when using supervised feature selection. It has been observed that an optimal selection of a large number of features is often an indication for overfitting. Kernel PCA with linear kernel gives better results.

Availability: Matlab scripts are available on request.

Contact: Nathalie.Pochet@esat.kuleuven.ac.be

Supplementary information: http://www.esat.kuleuven.ac.be/~npochet/Bioinformatics/

INTRODUCTION

Microarrays allow to determine the expression levels of thousands of genes simultaneously. One important application area of this technology is clinical oncology. As the dysregulated expression of genes lies at the origin of the tumor phenotype, its measurement can be very helpful to model or to predict the clinical behavior of malignancies. By these means, the fundamental processes underlying carcinogenesis can be integrated into the clinical decision making.

For clinical applications, microarray data can be represented by an expression matrix of which the rows represent the gene expression profiles and the columns the expression patterns of the patients. Using microarray data allows optimized predictions for an individual patient, e.g. predictions about therapy response, prognosis and metastatic phenotype. An example of the first one can be found in Iizuka *et al*. (2003). Hepatocellular carcinoma has a poor prognosis because of the high intrahepatic recurrence rate. Intrahepatic recurrence limits the potential of surgery as a cure for hepatocellular carcinoma. The current pathological prediction systems clinically applied to patients are inadequate for predicting recurrence in individuals who undergo hepatic resection. In this case, it would be useful to predict therapy response in order to be able to select the patients who would benefit from surgical treatment. An example of the second prediction is given in Nutt *et al*. (2003). Among high-grade gliomas, anaplastic oligodendrogliomas have a more favorable prognosis than glioblastomas. Moreover, although glioblastomas are resistant to most available therapies, anaplastic oligodendrogliomas are often chemosensitive. By predicting the prognosis, it is possible to finetune treatment. An example of the third prediction is presented in van't Veer *et al*. (2002). For breast cancer patients without tumor cells in local

*To whom correspondence should be addressed.

lymph nodes at diagnosis (lymph node negative), it is useful to predict the presence of distant subclinical metastases (poor prognosis) based on the primary tumor. Predicting the metastatic phenotype allows selecting patients who would benefit from adjuvant therapy as well as selecting patients for whom this adjuvant therapy would mean unnecessary toxicity.

Microarray datasets are characterized by high dimensionality in the sense of a small number of patients and a large number of gene expression levels for each patient. Most classification methods have problems with the high dimensionality of microarray data and require dimensionality reduction first. On the contrary, support vector machines (SVMs) are capable of learning and generalizing these data well (Mukherjee *et al*., 1999; Furey *et al*., 2000). Most classification methods like for example fisher discriminant analysis also rely on linear functions and are unable to discover non-linear relationships in microarray data, if any. By using kernel functions, one aims at better understanding of these data (Brown *et al*., 2000), especially when more patient data may become available in the future. The first aim of this study is to compare linear versions of the standard techniques applied to microarray data with their kernel version counterparts both with linear and radial basis function (RBF) kernel. Even with a linear kernel, least squares SVMs techniques can be more suitable as they contain regularization and do not require dimensionality reduction as applied in the dual space. A second aim is to find an optimal strategy for the performance of clinical predictions. In this paper, we systematically assess the role of dimensionality reduction and non-linearity on a wide variety of microarray datasets, instead of doing this in an *ad hoc* manner. Randomizations on all datasets are carried out in order to get a more reliable idea of the to be expected performance and the variation on it. The results on one specific partitioning of training, validation and test set (as often reported in literature) could easily lead to overly optimistic results, especially in the case of a small number of patient data.

SYSTEMATIC BENCHMARKING

Datasets

This study considers nine cancer classification problems, all comprising two classes. For this purpose, seven publically available microarray datasets are used: colon cancer data (Alon *et al*., 1999), acute leukemia data (Golub *et al*., 1999), breast cancer data (Hedenfalk *et al*., 2001), hepatocellular carcinoma data (Iizuka *et al*., 2003), high-grade glioma data (Nutt *et al*., 2003), prostate cancer data (Singh *et al*., 2002) and breast cancer data (van't Veer *et al*., 2002). Since the dataset in Hedenfalk *et al*. (2001) contains three classes, three binary classification problems and corresponding datasets can be constructed from it by taking each class versus the rest. In most of the datasets, all data samples have already been assigned to a training set or test set. In the cases of datasets for which a training set and test set have not been defined yet, two-third of the data samples of each class are assigned to the training set and the rest to the test set.

Table 1. Summary of the nine binary cancer classification problems datasets reflecting the dimensions and the microarray technology of each dataset

D	TR	TR C1	TR C2	TE	TE C1	TE C2	Levels	M
1	40	14	26	22	8	14	2000	T1
2	38	11	27	34	14	20	7129	T1
3	14	4	10	8	3	5	3226	T2
4	14	5	9	8	3	5	3226	T2
5	14	4	10	8	3	5	3226	T2
6	33	12	21	27	8	19	7129	T1
7	21	14	7	29	14	15	12625	T1
8	102	52	50	34	25	9	12600	T1
9	78	34	44	19	12	7	24188	T2

Explanation of the abbreviations used: D, datasets; TR, training set; TE, test set; C1, class 1; C2, class 2; M, microarray technology; T1, oligonucleotide; T2, cDNA; 1, colon cancer data of Alon *et al*. (1999); 2, acute leukemia data of Golub *et al*. (1999); 3, breast cancer data of Hedenfalk *et al*. (2001) taking the BRCA1 mutations versus the rest; 4, breast cancer data of Hedenfalk *et al*. (2001) taking the BRCA2 mutations versus the rest; 5, breast cancer data of Hedenfalk *et al*. (2001) taking the sporadic mutations versus the rest; 6, hepatocellular carcinoma data of Iizuka *et al*. (2003); 7, high-grade glioma data of Nutt *et al*. (2003); 8, prostate cancer data of Singh *et al*. (2002); and 9, breast cancer data of van't Veer *et al*. (2002).

An overview of the characteristics of all the datasets can be found in Table 1. The acute leukemia data in Golub *et al*. (1999) have already been used frequently in previous microarray data analysis studies. Preprocessing of this dataset is done by thresholding and log-transformation, similar as in the original publication. Thresholding is achieved by restricting gene expression levels to be larger than 20, e.g. expression levels which are smaller than 20 will be set to 20. Concerning the log-transformation, the natural logarithm of the expression levels is taken. The breast cancer dataset in van't Veer *et al*. (2002) contains missing values. Those have been estimated based on 5% of the gene expression profiles that have the largest correlation with the gene expression profile of the missing value. No further preprocessing is applied to the rest of the datasets.

Systematic benchmarking studies are important for obtaining reliable results allowing comparability and repeatability of the different numerical experiments. For this purpose, this study not only uses the original division of each dataset in training and test set, but also reshuffles (randomizes) all datasets. Consequently, all numerical experiments are performed with 20 randomizations of the 9 original datasets as well. These randomizations are the same for all numerical experiments on one dataset (in Matlab with the same seed for the random generator). They are also stratified, which means that each randomized training and test set contains the same amount of samples of each class compared to the original training and test set. The results of all numerical experiments in the tables represent the mean and SD of the results on each original dataset and 20 randomizations.

Methods

The methods used to set up the numerical experiments can be subdivided in two categories: dimensionality reduction and classification. For dimensionality reduction, classical principal component analysis (PCA) as well as kernel PCA are used. Fisher discriminant analysis (FDA) and LS-SVM (which can be viewed among others as a kernel version of FDA) are used for classification.

Principal component analysis PCA looks for linear combinations of gene expression levels in order to obtain a maximal variance over a set of patients. In fact, those combinations are most informative for this set of patients and are called the principal components. One formulation in order to characterize PCA problems is to consider a given set of centered (zero mean) input data $\{x_k\}_{k=1}^N$ as a cloud of points for which one tries to find projected variables $w^T x$ with maximal variance. This means,

$$\max_w \operatorname{Var}(w^T x) = w^T C w, \tag{1}$$

where the covariance matrix C is estimated as $C \cong (1/(N-1)) \sum_{k=1}^N x_k x_k^T$. One optimizes this objective function under the constraint that $w^T w = 1$. Solving the constrained optimization problem gives the eigenvalue problem

$$Cw = \lambda w. \tag{2}$$

The matrix C is symmetric and positive semidefinite. The eigenvector w corresponding to the largest eigenvalue determines the projected variable having maximal variance.

Kernel principal component analysis Kernel PCA has the same goal as classical PCA, but is capable of looking for non-linear combinations too. The objective of kernel PCA can be formulated (Schölkopf *et al.*, 1998; Suykens *et al.*, 2003) as

$$\max_w \sum_{k=1}^N [w^T(\varphi(x_k) - \mu_\varphi)]^2, \tag{3}$$

with notation $\mu_\varphi = (1/N)\sum_{k=1}^N \varphi(x_k)$ used for centering the data in the feature space, where $\varphi(\cdot)\colon \mathbb{R}^n \to \mathbb{R}^{n_h}$ is the mapping to a high-dimensional feature space, which might be infinite dimensional. This can be interpreted as first mapping the input data to a high dimensional feature space and next to projected variables. The following optimization problem is formulated in the primal weight space

$$\max_{w,e} J_P(w,e) = \gamma \frac{1}{2}\sum_{k=1}^N e_k^2 - \frac{1}{2} w^T w,$$

$$\text{such that } e_k = w^T[\varphi(x_k) - \mu_\varphi], \quad k = 1, \ldots, N. \tag{4}$$

This formulation states that the variance of the projected variables is maximized for the given N data points while keeping the norm of w small by the regularization term. By taking the conditions for optimality from the Lagrangian related to this constrained optimization problem, such as $w = \sum_{k=1}^N \alpha_k[\varphi(x_k) - \mu_\varphi]$ among others, and defining $\lambda = 1/\gamma$, one obtains the eigenvalue problem

$$\Omega_c \alpha = \lambda \alpha, \tag{5}$$

with

$$\Omega_{c,kl} = [\varphi(x_k) - \mu_\varphi]^T[\varphi(x_l) - \mu_\varphi], \quad k, l = 1, \ldots, N, \tag{6}$$

the elements for the centered kernel matrix Ω_c. Since the kernel trick $K(x_k, x_l) = \varphi(x_k)^T \varphi(x_l)$ can be applied to the centered kernel matrix, one may choose any positive definite kernel satisfying the Mercer condition. The kernel functions used in this paper are the linear kernel $K(x, x_k) = x_k^T x$ and the RBF kernel $K(x, x_k) = \exp\{-\|x - x_k\|_2^2/\sigma^2\}$. The centered kernel matrix can be computed as $\Omega_c = M_c \Omega M_c$ with $\Omega_{kl} = K(x_k, x_l)$ and $M_c = I - (1/N)1_N 1_N^T$ the centering matrix where I denotes the identity matrix and 1_N is a vector of length N containing all ones. The dimensionality reduction is done by selecting the eigenvectors corresponding to the largest eigenvalues.

Fisher discriminant analysis FDA projects the data $x_k \in \mathbb{R}^n$ from the original input space to a one-dimensional variable $z_k \in \mathbb{R}$ and makes a discrimination based on this projected variable. In this one-dimensional space one tries to achieve a high discriminatory power by maximizing the between-class variances and to minimize the within-class variances for the two classes. The data are projected as follows

$$z = f(x) = w^T x + b, \tag{7}$$

with $f(\cdot)\colon \mathbb{R}^n \to \mathbb{R}$. One is interested then in finding a line such that the following objective of a Rayleigh quotient is maximized:

$$\max_{w,b} J_{FD}(w,b) = \frac{w^T \Sigma_B w}{w^T \Sigma_W w}. \tag{8}$$

The means of the input variables for class 1 and class 2 are $\mathcal{E}[x^{(1)}] = \mu^{(1)}$, $\mathcal{E}[x^{(2)}] = \mu^{(2)}$. The between and within covariance matrices related to class 1 and class 2 are $\Sigma_B = [\mu^{(1)} - \mu^{(2)}][\mu^{(1)} - \mu^{(2)}]^T$, $\Sigma_W = \mathcal{E}\{[x - \mu^{(1)}][x - \mu^{(1)}]^T\} + \mathcal{E}\{[x - \mu^{(2)}][x - \mu^{(2)}]^T\}$ where the latter is the sum of the two covariance matrices Σ_{W_1}, Σ_{W_2} for the two classes. Note that the Rayleigh quotient is independent of the bias term b. By choosing a threshold z_0, it is possible to classify a new point as belonging to class 1 if $z(x) \geq z_0$, and classify it as belonging to class 2 otherwise. Assuming that the projected data is the sum of a set of random variables allows invoking the central limit theorem and modeling the class-conditional density functions $p(z|$ class 1) and $p(z|$ class 2) using normal distributions.

Least squares support vector machine classifiers LS-SVMs (Suykens and Vandewalle, 1999; Van Gestel *et al*., 2002; Pelckmans *et al*., 2002, http://www.esat.kuleuven.ac.be/sista/lssvmlab/) are a modified version of SVMs (Vapnik, 1998; Schölkopf *et al*., 1999, 2001; Cristianini and Shawe-Taylor, 2000; Schölkopf and Smola, 2002) and comprises a class of kernel machines with primal-dual interpretations related to kernel FDA, kernel PCA, kernel PLS (kernel Partial Least Squares), kernel CCA (kernel Canonical Correlation Analysis), recurrent networks and others. For classification this modification leads to solving a linear system instead of a quadratic programming problem, which makes LS-SVM much faster than SVM on microarray datasets. The benchmarking study of Van Gestel *et al*. (2004) on 20 UCI datasets revealed that the results of LS-SVM are similar to those of SVM. Given is a training set $\{x_k, y_k\}_{k=1}^N$ with input data $x_k \in \mathbb{R}^n$ and corresponding binary class labels $y_k \in \{-1, +1\}$. Vapnik's SVM classifier formulation was modified in (Suykens and Vandewalle, 1999) into the following LS-SVM formulation:

$$\begin{gathered}\min_{w,b,e} \; J_P(w,e) = \tfrac{1}{2} w^T w + \gamma \tfrac{1}{2} \textstyle\sum_{k=1}^N e_k^2, \\ \text{such that } y_k[w^T \varphi(x_k) + b] = 1 - e_k, \quad k = 1, \ldots, N,\end{gathered} \tag{9}$$

for a classifier in the primal space that takes the form

$$y(x) = \operatorname{sign}[w^T \varphi(x) + b], \tag{10}$$

where $\varphi(\cdot)\colon \mathbb{R}^n \to \mathbb{R}^{n_h}$ is the mapping to the high-dimensional feature space and γ the regularization parameter. In the case of a linear classifier one could easily solve the primal problem, but in general w might be infinite dimensional. For this nonlinear classifier formulation, the Lagrangian is solved, which results in the following dual problem to be solved in α, b:

$$\left[\begin{array}{c|c} 0 & y^T \\ \hline y & \Omega + I/\gamma \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline 1_N \end{array}\right], \tag{11}$$

where the kernel trick $K(x_k, x_l) = \varphi(x_k)^T \varphi(x_l)$ can be applied within the Ω matrix

$$\Omega_{kl} = y_k y_l \varphi(x_k)^T \varphi(x_l) = y_k y_l K(x_k, x_l), \quad k, l = 1, \ldots, N. \tag{12}$$

The classifier in the dual space takes the form

$$y(x) = \sum_{k=1}^N \alpha_k y_k K(x, x_k) + b. \tag{13}$$

The chosen kernel function should be positive definite and satisfy the Mercer condition. The kernel functions used in this paper are the linear kernel $K(x, x_k) = x_k^T x$ and the RBF kernel $K(x, x_k) = \exp\{-\|x - x_k\|_2^2/\sigma^2\}$. Note that using LS-SVM with a linear kernel without regularization ($\gamma \to \infty$) is in fact the counterpart of classical linear FDA, but the latter needs dimensionality reduction while the former can handle the problem without dimensionality reduction in the dual form as the size of the linear system to be solved is $(N+1) \times (N+1)$ and is not determined by the number of gene expression levels. Hence, the advantage of using kernel methods like SVM or LS-SVM is that they can be used without performing dimensionality reduction first, which is not the case for the classical linear regression method FDA.

Numerical experiments

In this study, nine classification problems are considered. The numerical experiments applied to all these problems can be divided into two subgroups, depending on the required parameter optimization procedure. First, three kinds of experiments, all without dimensionality reduction, are performed to all nine classification problems. These are LS-SVM with linear kernel, LS-SVM with RBF kernel and LS-SVM with linear kernel and infinite regularization parameter ($\gamma \to \infty$). Next, six kinds of experiments, all using dimensionality reduction, are performed to all nine classification problems. The first two of these are based on classical PCA followed by FDA. Selection of the principal components is done both in an unsupervised and a supervised way. The same strategy is used in the last four of these, but kernel PCA with linear kernel as well as RBF kernel are used instead of classical linear PCA.

Since building a prediction model requires good generalization towards making predictions for previously unseen test samples, tuning the parameters is an important issue. The small sample size characterizing microarray data restricts the choice of an estimator for the generalization performance. The optimization criterion used in this study is the leave-one-out cross-validation (LOO-CV) performance. In each LOO-CV iteration (number of iterations equals the sample size), one sample is left out of the data, a classification model is trained on the rest of the data and this model is then evaluated on the left out data point. As an evaluation measure, the LOO-CV performance $[(\text{No. of correctly classified samples})/(\text{No. of samples in the data}) \cdot 100]\%$ is used.

All numerical experiments are implemented in Matlab by using the LS-SVM and kernel PCA implementations of the LS-SVMlab toolbox (http://www.esat.kuleuven.ac.be/sista/lssvmlab/).

Tuning parameter optimization for the case without dimensionality reduction When using LS-SVM with a linear kernel, only the regularization constant needs to be further optimized. The value of the regularization parameter corresponding to the largest LOO-CV performance is then selected as the optimal value. Using an RBF kernel instead requires optimization of the regularization parameter γ as well as the kernel parameter σ. This is done by searching a two dimensional grid of different values for both parameters. Using LS-SVM with a linear kernel and infinite regularization parameter, which corresponds to FDA, requires no parameter optimization.

After preprocessing, which is specific for each dataset (as discussed in the section on datasets), normalization is always performed on all the datasets before using them for classification purposes. This is done by standardizing each gene expression of the data to have zero mean and unit SD. Normalization of training sets as well as test sets is done by using the mean and SD of each gene expression profile of the training sets.

Tuning parameter optimization in the case of dimensionality reduction When reducing the dimensionality of the expression patterns of the patients with classical PCA and next building a prediction model by means of FDA, the number of principal components needs to be optimized. This is realized by performing LOO-CV on the training set. For each possible number of principal components (ranging between 1 and $N - 2$, with N the number of training samples), the LOO-CV performance is computed. The number of principal components with best LOO-CV performance is then selected as the optimal one. If there exist different numbers of principal components with the same best LOO-CV performance, the smallest number of principal components is selected. This choice can be interpreted as minimizing the complexity of the model. In case kernel PCA with a linear kernel is used instead of the classical PCA, the same method is used. Using kernel PCA with an RBF kernel not only requires optimization of the number of principal components, but also the kernel parameter σ needs to be tuned. A broad outline of the optimization procedure is described in the sequel. For several possible values of the kernel parameter, the LOO-CV performance is computed for each possible number of principal components. The optimal number of principal components with the best LOO-CV performance is then selected for each value of the kernel parameter. If there are several optimal numbers of principal components, the smallest number of principal components is selected, again for minimal model complexity reasons. In order to find the optimal value for the kernel parameter, the value of the kernel parameter with best LOO-CV performance is selected. In case there are several possible optimal values for the kernel parameter, also the optimal number of principal components belonging to these optimal kernel parameter values need to be considered. From these values, the optimal kernel parameter value with the smallest number of principal components is chosen. In case there are still several possible optimal kernel parameter values, the smallest value of these is selected as the optimal one. Remark the complexity of this optimization procedure because both the kernel parameter and the number of principal components of the kernel PCA with RBF kernel need to be optimized in the sense of the LOO-CV performance of the FDA classification.

Optimization algorithm: kernel PCA with RBF kernel followed by FDA

```
(1) Generation of parameter grid
    for each kernel parameter value within selected range
        for each possible # principal components
            for each LOO-CV iteration
                • leave one sample out
                • normalization
                • dimensionality reduction (kernel PCA)
                • selection of the principal components (un
                  supervised or supervised)
                • classification (FDA)
                • test sample left out
            end
            calculate LOO-CV performance
        end
    end
(2) Optimization of parameters
    for each kernel parameter value out of a range
        optimal # principal components:
          1. best LOO-CV performance
          2. smallest # principal components *
    end
    optimal kernel parameter value:
       1. best LOO-CV performance
       2. smallest # principal components *
       3. smallest kernel parameter value *
* if more than one
```

Normalization of the samples left out in each LOO-CV iteration also needs to be done based on the mean and SD of each gene expression profile of each accompanying training set. Concerning dimensionality reduction, it should be remarked that this is also done based on the training set. First, PCA is applied to the training set, which results in eigenvalues and eigenvectors going from 1 till N. The training and test set are then projected onto those eigenvectors. As the data are centered, the last eigenvalue is equal to zero. Therefore, the last principal component is left out, which results in the number of principal components going from 1 till $N - 2$. In fact, this corresponds to obtaining a low-rank approximation starting from a full rank matrix.

Supervised versus unsupervised selection of principal components Concerning the experiments with dimensionality reduction, two ways of selecting the principal components are used. The first one simply looks at the eigenvalues of the principal components, originating from PCA. Since this method does not take into account the class labels, it is in an unsupervised way. The other one is based on the absolute value of the score introduced by Golub *et al*. (1999), as also used in Furey *et al*. (2000):

$$F(x_j) = \left| \frac{\mu_j^1 - \mu_j^2}{\sigma_j^1 + \sigma_j^2} \right|. \tag{14}$$

This method allows finding individual gene expression profiles that help discriminating between two classes by calculating for each gene expression profile x_j a score based on the mean μ_j^1 (respectively μ_j^2) and the SD σ_j^1 (respectively σ_j^2) of each class of samples. In our experiments, this method is applied onto the principal components instead of applying it directly to the gene expression profiles. This method takes into account the class labels and is therefore called supervised. The n most important principal components now correspond to the n principal components with either the highest eigenvalues or the highest absolute value of the score introduced by Golub.

Measuring and comparing the performance of the numerical experiments For the results, three kinds of measures are used. The first one is the LOO-CV performance. This is estimated by only making use of the training datasets for tuning the parameters. The second measure is the accuracy, which gives an idea of the classification performance by reflecting the percentage correctly classified samples. When measured on independent test sets, this gives an idea of the generalization performance. But when measured on the training set, one can get an idea of the degree of overfitting. The third measure is the area under the Receiver operating characteristic (ROC) curve (Hanley and McNeil, 1982). An ROC curve shows the separation abilities of a binary classifier: by setting different possible classifier thresholds, the performances [(No. of correctly classified samples)/(No. of samples in the data) · 100]% are calculated resulting in the ROC curve. If the area under the ROC curve equals 100% on a dataset, a perfectly separating classifier is found on that particular dataset, if the area equals 50%, the classifier has no discriminative power at all. This measure can be evaluated on an independent test set or training set. Statistical significance tests are performed in order to allow a correct interpretation of the results. A non-parametric paired test, the Wilcoxon signed rank test (signrank in Matlab) (Dawson-Saunders and Trapp, 1994), has been used in order to make general conclusions. A threshold of 0.05 is respected, which means that two results are statistically significantly different if the value of the Wilcoxon signed rank test applied to both of them is lower than 0.05.

RESULTS

The tables with all results and the statistical significance tests as well as a detailed description of all nine classification problems can be found on the supplementary website. Only the most relevant classification problems are treated in the following discussion and are represented in Table 2. For each classification problem, the results represent the statistical summary (mean and variance) of the numerical experiments on the original dataset and 20 randomizations of it. Since the randomizations (training and test set splits) are not disjoint, the results as well as the statistical significance tests given in the tables are not unbiased and can in general also be too optimistic.

General comments

One general remark is that constructing the randomizations in a stratified way already seems to result in a large variance (it would have been even larger if constructed in a non-stratified way).

Another remark is that the LOO-CV performance is not a good indicator for the accuracy or the area under the ROC curve of the test set. This raises the question whether or not this LOO-CV performance is a good method for tuning the parameters. Since microarray data are characterized by a small sample size, LOO-CV has to be applied with care as one may easily overfit in this case.

For all datasets except the one containing the acute leukemia data (Golub *et al*., 1999), the LOO-CV performance, the test set accuracy and also the area under the ROC curve of the test set of the experiment based on LS-SVM with linear kernel and $\gamma \to \infty$ (i.e. no regularization) is significantly worse than all other experiments. This clearly indicates that regularization is very important when performing classification without previous dimensionality reduction, even for linear models. In the further discussion treating the individual datasets, this experiment will be left out.

The acute leukemia data (Golub *et al*., 1999) clearly comprises an easy classification problem, since the variances on the results caused by the randomizations are quite small compared to the other datasets. All experiments on this dataset also seem to end up in quite similar results, so in fact it hardly doesn't matter which classification method is applied on this dataset.

Observing the optimal values for the tuning parameters leads to the following remarks. When LS-SVM with a linear kernel is applied, typical values for the mean regularization parameter γ on each dataset are ranging between 1e−3 and 1e+3. When using LS-SVM with an RBF kernel, typical values for the mean regularization parameter γ as well as the mean kernel parameter σ^2 on each dataset both are ranging between 1e+10 and 1e+15. Optimal values for the kernel parameter σ^2 are quite large because they are scaled with the large input dimensionality of microarray data. Using kernel PCA with an RBF kernel before classification often results in test set performances that are worse than when using kernel PCA with a linear kernel, which means that overfitting occurs. Typical values for the mean kernel parameter σ^2 of the kernel PCA with RBF kernel on each dataset highly

Table 2. Summary of the results of the numerical experiments on four binary cancer classification problems, comprising the LOO-CV performance, the accuracy (ACC) on training and test set, and the area under the ROC curve (AUC) on training and test set

	LOO-CV performance	ACC training set	ACC test set	AUC training set	AUC test set
Hedenfalk *et al*. (2001) BRCA1 mutations					
LS-SVM linear kernel	78.23 ± 7.13	87.76 ± 14.14	64.29 ± 6.99	100.00 ± 0.00	**81.90 ± 18.19** (+)
LS-SVM RBF kernel	82.65 ± 8.12	98.64 ± 6.08	**75.00 ± 12.20** (+)	100.00 ± 0.00	**82.22 ± 17.38** (+)
LS-SVM linear kernel (no regularization)	46.94 ± 21.21	47.62 ± 9.94	**52.98 ± 19.25** (−)	47.14 ± 14.38	**52.70 ± 24.16** (−)
PCA + FDA (unsupervised PC selection)	81.63 ± 7.17	95.24 ± 7.09	64.29 ± 12.96	93.93 ± 12.67	67.62 ± 21.83
PCA + FDA (supervised PC selection)	84.01 ± 9.58	97.96 ± 4.49	68.45 ± 15.25	97.86 ± 5.25	71.75 ± 21.12
kPCA lin + FDA (unsupervised PC selection)	81.29 ± 7.13	95.24 ± 6.73	63.10 ± 13.07	96.55 ± 5.64	66.35 ± 20.23
kPCA lin + FDA (supervised PC selection)	84.35 ± 8.99	98.30 ± 4.36	67.86 ± 15.70	98.45 ± 4.12	72.38 ± 22.23
kPCA RBF + FDA (unsupervised PC selection)	91.16 ± 7.28	94.90 ± 6.29	**54.17 ± 11.79** (−)	95.36 ± 7.98	60.63 ± 16.25
kPCA RBF + FDA (supervised PC selection)	92.52 ± 5.16	98.30 ± 5.36	63.69 ± 10.85	97.68 ± 7.72	64.13 ± 18.54
Nutt *et al*. (2003)					
LS-SVM linear kernel	75.74 ± 8.93	90.02 ± 14.16	61.25 ± 11.75	99.47 ± 1.03	79.25 ± 6.06
LS-SVM RBF kernel	78.23 ± 7.99	98.41 ± 7.10	**69.95 ± 8.59** (+)	100.00 ± 0.00	**81.04 ± 6.64** (+)
LS-SVM linear kernel (no regularization)	50.79 ± 16.65	50.79 ± 12.75	**48.93 ± 10.88** (−)	50.63 ± 16.40	**50.68 ± 15.15** (−)
PCA + FDA (unsupervised PC selection)	80.95 ± 7.49	92.29 ± 7.12	67.82 ± 7.24	97.72 ± 2.80	77.48 ± 10.50
PCA + FDA (supervised PC selection)	81.41 ± 7.19	92.97 ± 10.14	65.52 ± 11.01	96.65 ± 5.69	77.37 ± 9.04
kPCA lin + FDA (unsupervised PC selection)	80.73 ± 7.12	92.52 ± 6.98	68.31 ± 6.78	97.91 ± 2.74	77.98 ± 10.43
kPCA lin + FDA (supervised PC selection)	81.86 ± 6.67	95.24 ± 8.57	67.32 ± 11.04	98.15 ± 4.02	76.53 ± 8.96
kPCA RBF + FDA (unsupervised PC selection)	86.62 ± 5.99	94.78 ± 9.05	**64.20 ± 11.19** (−)	97.30 ± 6.60	**70.80 ± 15.44** (−)
kPCA RBF + FDA (supervised PC selection)	85.94 ± 5.78	96.15 ± 7.29	**58.13 ± 12.24** (−)	98.25 ± 3.78	**66.33 ± 15.48** (−)
Singh *et al*. (2002)					
LS-SVM linear kernel	90.10 ± 1.42	100.00 ± 0.00	84.31 ± 13.66	100.00 ± 0.00	**91.28 ± 5.20** (+)
LS-SVM RBF kernel	91.22 ± 1.19	99.95 ± 0.21	**88.10 ± 4.93** (+)	100.00 ± 0.00	**92.04 ± 5.03** (+)
LS-SVM linear kernel (no regularization)	50.33 ± 0.92	51.45 ± 7.03	**48.18 ± 10.25** (−)	51.10 ± 8.27	**50.98 ± 12.38** (−)
PCA + FDA (unsupervised PC selection)	90.38 ± 1.83	97.62 ± 1.95	83.89 ± 13.63	99.67 ± 0.38	88.93 ± 11.39
PCA + FDA (supervised PC selection)	90.57 ± 1.53	97.57 ± 3.34	82.49 ± 13.35	99.40 ± 0.99	86.74 ± 12.95
kPCA lin + FDA (unsupervised PC selection)	90.34 ± 1.75	97.57 ± 1.90	85.01 ± 9.07	99.67 ± 0.38	89.98 ± 7.30
kPCA lin + FDA (supervised PC selection)	90.57 ± 1.53	97.57 ± 3.34	82.49 ± 13.35	99.40 ± 0.99	86.73 ± 12.96
kPCA RBF + FDA (unsupervised PC selection)	91.60 ± 1.50	98.97 ± 1.75	85.01 ± 11.00	99.84 ± 0.32	89.90 ± 9.64
kPCA RBF + FDA (supervised PC selection)	100.00 ± 0.00	100.00 ± 0.00	**28.71 ± 10.02** (−)	100.00 ± 0.00	**50.00 ± 0.00** (−)
Van ’t Veer *et al*. (2002)					
LS-SVM linear kernel	68.99 ± 4.22	100.00 ± 0.00	**67.92 ± 8.58** (+)	100.00 ± 0.00	**73.30 ± 11.01** (+)
LS-SVM RBF kernel	69.05 ± 3.55	100.00 ± 0.00	**68.42 ± 7.62** (+)	100.00 ± 0.00	**73.98 ± 10.69** (+)
LS-SVM linear kernel (no regularization)	52.14 ± 6.04	74.66 ± 24.04	**57.14 ± 9.08** (−)	74.73 ± 25.26	**64.60 ± 13.18** (−)
PCA + FDA (unsupervised PC selection)	71.31 ± 3.57	91.27 ± 10.04	57.39 ± 15.57	94.61 ± 6.80	65.16 ± 12.30
PCA + FDA (supervised PC selection)	73.44 ± 3.19	97.31 ± 5.62	**66.92 ± 9.90** (+)	98.77 ± 3.16	67.91 ± 12.64
kPCA lin + FDA (unsupervised PC selection)	71.18 ± 3.62	91.21 ± 10.33	60.90 ± 14.49	94.46 ± 7.22	66.01 ± 13.45
kPCA lin + FDA (supervised PC selection)	73.63 ± 3.89	97.13 ± 6.63	**65.41 ± 7.54** (+)	98.54 ± 3.98	69.22 ± 11.01
kPCA RBF + FDA (unsupervised PC selection)	74.91 ± 6.54	90.66 ± 11.08	51.38 ± 15.91	93.77 ± 8.75	60.26 ± 16.57
kPCA RBF + FDA (supervised PC selection)	100.00 ± 0.00	100.00 ± 0.00	**36.84 ± 0.00** (−)	100.00 ± 0.00	**50.00 ± 0.00** (−)

The results visualized in bold followed by (+) are statistically significantly better than the other results. The results in bold followed by (−) are statistically significantly worse than the other results.

depend on the way the principal components are selected. When using the unsupervised way for selecting the principal components, the mean of kernel parameter values σ^2 tends to go to 1e+20. Using the supervised way for selecting the principal components, $1\mathrm{e}+0$ is often selected as the optimal value for the kernel parameter σ^2, which leads to bad test set performances compared to the other experiments (seriously overfitting).

In the context of parameter optimization, it is also important to address the number of selected features and in particular the sparseness of the classical and kernel PCA projections. Figure 1 represents the test set ROC performance together with the sparseness when using a linear and an RBF kernel for kernel PCA. It has been noticed that classical PCA leads to approximately the same results as kernel PCA with linear kernel and therefore not represented separately. Selection of

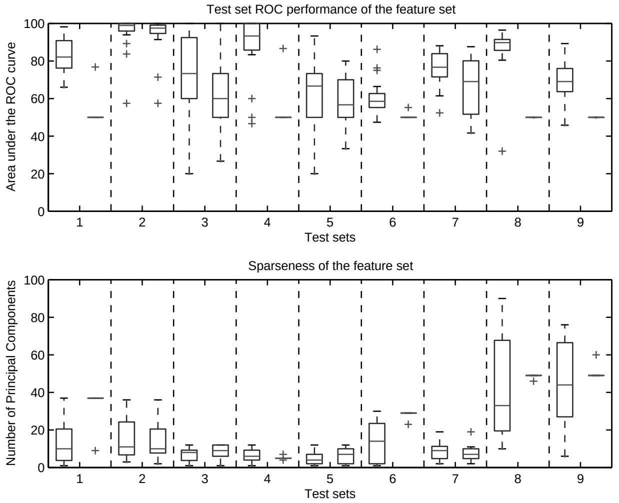

Fig. 1. Illustration of the test set ROC performance (upper part) and the sparseness (lower part) of the optimally selected feature set based on boxplots of the areas under the ROC curve of the test set and boxplots of the optimal number of principal components respectively of all nine cancer classification problems. It has been observed that an optimal selection of a large number of features is often an indication for overfitting in case of kernel PCA with RBF kernel (supervised feature selection) followed by FDA. For each dataset, the areas under the ROC curve of the test set and the optimal number of principal components of kernel PCA with a linear kernel (selecting the principal components in a supervised way) followed by FDA are represented on the left, the areas under the ROC curve of the test set and the optimal number of principal components of kernel PCA with an RBF kernel (selecting the principal components in a supervised way) followed by FDA on the right. Concerning the datasets, the order of Table 1 is respected.

the principal components is done in a supervised way based on the LOO-CV performance criterion. Two observations can be stated when comparing the results of these two experiments. First, when the optimal number of principal components is relatively low in case of using a linear kernel and much larger in case of using an RBF kernel, this is an indication of overfitting. The colon cancer dataset of (Alon *et al.*, 1999) (1) and the hepatocellular carcinoma dataset of (Iizuka *et al.*, 2003) (6) are examples of this observation. Second, when the optimal number of principal components is very large both in case of using a linear kernel and in case of using an RBF kernel, this is an indication of overfitting too. The prostate cancer dataset of (Singh *et al.*, 2002) (8) and the breast cancer dataset of (van't Veer *et al.*, 2002) (9) are illustrating this observation.

Results on specific datasets

Breast cancer dataset (Hedenfalk et al., 2001): BRCA1 mutations versus the rest. Concerning the test set accuracies, LS-SVM with RBF kernel obviously performs better than all other methods. Using an RBF kernel when doing kernel PCA on the other hand clearly performs worse when the eigenvalues are used for selection of the principal components. The results of the area under the ROC curve of the test set show that using LS-SVM results in much better performances than all other experiments, even when using a linear kernel. Both methods for selecting the principal components seem to perform very similarly, but in some cases using the absolute value of the Golub score tends to perform slightly better. Remarkably in this case is that the test set accuracy of LS-SVM with RBF kernel is much better than LS-SVM with linear kernel, although the area under the ROC curve of both experiments is practically equal. This is also an indication of how important it is to find a good decision threshold value, which corresponds to an operating point on the ROC curve.

High-grade glioma dataset (Nutt et al., 2003). Concerning the test set performances, the experiment using LS-SVM with RBF kernel is significantly better than using LS-SVM with linear kernel. For this dataset both methods for selection of the principal components give similar results.

Prostate cancer dataset (Singh et al., 2002). The test set performances show that the experiment using kernel PCA with RBF kernel and selecting the principal components by means of the supervised method clearly gives very bad results. Using the eigenvalues for selection of the principal components seems to give better results than using the supervised method. According to the test set accuracy, the experiment applying LS-SVM with RBF kernel even performs slightly better than those experiments using the eigenvalues for selection of the principal components. When looking at the area under the ROC curve of the test set, both experiments applying LS-SVM perform slightly better than those experiments using the eigenvalues for selection of the principal components.

Breast cancer dataset (van't Veer et al., 2002). When looking at the test set performances, it is obvious that the experiment using kernel PCA with RBF kernel and selecting the principal components by means of the supervised method leads to very bad results. Using LS-SVM gives better results than performing dimensionality reduction combined with an unsupervised way for the selection of the principal components. According to the area under the ROC curve of the test set, using LS-SVM gives better results than all experiments performing dimensionality reduction. Both methods for selecting the principal components seem to perform very similarly, but in some cases using the absolute value of the Golub score tends to perform slightly better.

DISCUSSION

Assessing the role of non-linearity for the case without dimensionality reduction

When considering only the experiments without dimensionality reduction, i.e. LS-SVM with linear kernel and LS-SVM

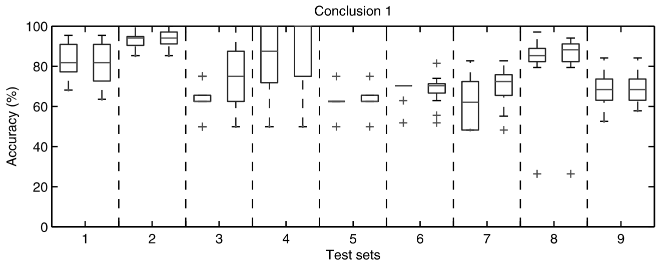

Fig. 2. Illustration of the first main conclusion based on boxplots (boxplot in Matlab, see supplementary website) of the test set accuracies of all nine binary cancer classification problems: When performing classification with LS-SVM (without dimensionality reduction), using well-tuned RBF kernels can be applied without risking overfitting. The results obtained with well-tuned RBF kernels are never worse and sometimes even statistically significantly better compared with using a linear kernel. For each dataset, the test set accuracies of LS-SVM with a linear kernel are represented on the left, the test set accuracies of LS-SVM with an RBF kernel on the right. Concerning the datasets, the order of Table 1 is respected.

with RBF kernel, using a well-tuned RBF kernel never resulted in overfitting on all tried datasets. The test set performances obtained when using an RBF kernel often appear to be similar to those obtained when using a linear kernel, but in some cases an RBF kernel ends up in even better classification performances. This is illustrated in Figure 2. The fact that using LS-SVM with an RBF kernel does not result in overfitting even for simple classification problems, can be explained by looking to the optimal values of the kernel parameter. When optimizing the kernel parameter of the RBF kernel for such a problem, the obtained value seems to be very large. Using an RBF kernel with the kernel parameter σ set to infinity corresponds to using a linear kernel, aside from a scale factor (Suykens *et al*., 2002). Until now, most microarray datasets are quite small and they may represent quite easily separable classification problems. It can be expected that those datasets will become larger or perhaps represent more complex classification problems in the future. In this case the use of non-linear kernels as the commonly used RBF kernel becomes important. Considering this, it may be useful to explore the effect of using other kernel functions.

When comparing the experiments with and without dimensionality reduction, an important issue is that LS-SVM with RBF kernel (experiment without dimensionality reduction) never performs worse than all other methods.

The importance of regularization

When looking at the experiment using LS-SVM with linear kernel and the regularization parameter γ set to infinity, i.e. without regularization, the following issue can be seen. Using LS-SVM without regularization corresponds to FDA (Suykens *et al*., 2002). Figure 3 shows that this experiment hardly performs better than random classification on all datasets, except on the acute leukemia dataset of (Golub *et al*., 1999), which represents an easily separable classification problem. Regularization appears to be very important when applying classification methods onto microarray data without doing a dimensionality reduction step first.

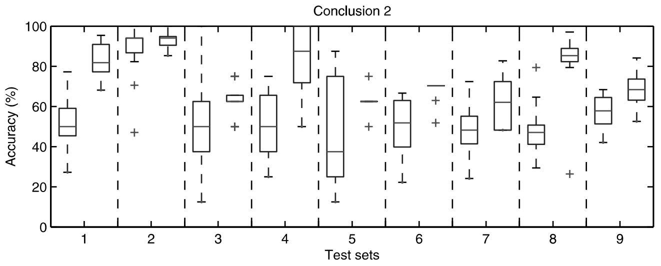

Fig. 3. Illustration of the second main conclusion based on boxplots of the test set accuracies of all nine cancer classification problems: Even for classification with linear classifiers like LS-SVM with linear kernel, performing regularization is very important. For each dataset, the test set accuracies of LS-SVM with a linear kernel without regularization are represented on the left, the test set accuracies of LS-SVM with a linear kernel with regularization on the right. The latter shows much better performance. Concerning the datasets, the order of Table 1 is respected.

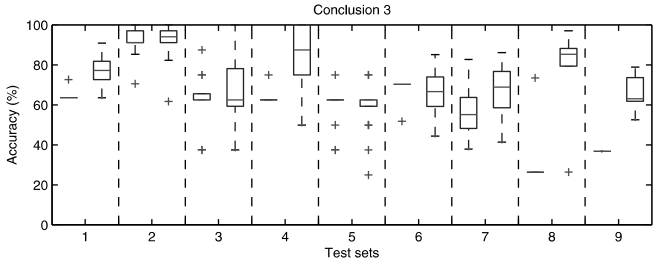

Fig. 4. Illustration of the third main conclusion based on boxplots of the test set accuracies of all nine cancer classification problems: When performing kernel principal component analysis (kernel PCA) before classification, using an RBF kernel for kernel PCA tends to result in overfitting. Kernel PCA with linear kernel gives better results. For each dataset, the test set accuracies of kernel PCA with an RBF kernel (selecting the principal components in a supervised way) followed by FDA are represented on the left, the test set accuracies of kernel PCA with a linear kernel (selecting the principal components in a supervised way) followed by FDA on the right. Concerning the datasets, the order of Table 1 is respected.

Assessing the role of non-linearity in case of dimensionality reduction

When considering only the experiments using dimensionality reduction, another important issue becomes clear. Comparing the results of using an RBF kernel with those of using a linear kernel when applying kernel PCA before classification, reveals that using an RBF kernel easily results in overfitting. This is represented by Figure 4. The best results are obtained by simply using a linear kernel when doing kernel PCA, which are similar to those when using classical PCA. (Gupta *et al*., 2002) states a similar conclusion for face recognition based on image data. When comparing both methods for selection of the principal components, namely the unsupervised way based on the eigenvalues with the supervised way based on the absolute value of the score introduced by (Golub *et al*., 1999), no general conclusions can be made. It depends on the dataset whether one method is better than the other or not. The combination of using kernel PCA with RBF kernel and selection of the principal components tends to result in overfitting. All this can be explained by ignoring relevant principal components (Bishop, 1995).

In the context of feature selection, some interesting issues become clear when studying the ROC performance and the sparseness of the classical and kernel PCA projections. When comparing the results of using a linear kernel with those of using an RBF kernel for kernel PCA when selection of the principal components is done in a supervised way as shown in Figure 1, two situations indicating overfitting can be recognized. First, overfitting occurs when the optimal number of principal components is relatively low in case of using a linear kernel for kernel PCA and much larger in case of using an RBF kernel. Second, overfitting also occurs when the optimal number of principal components is very large both in case of using a linear kernel for kernel PCA and in case of using an RBF kernel.

When comparing the experiments with and without dimensionality reduction, also worth mentioning is the fact that performing dimensionality reduction requires optimization of the number of principal components. This parameter, belonging to the unsupervised PCA, needs to be optimized in the sense of the subsequent supervised FDA (see outline of the optimization algorithm in the section on numerical experiments). In practice, this appears to be quite time-consuming, especially in combination with other parameters that need to be optimized (e.g. kernel parameter of kernel PCA with RBF kernel). However, numerical techniques can be used to speed up the experiments.

CONCLUSION

In the past, using classification methods in combination with microarrays has shown to be promising for guiding clinical management in oncology. In this study, several important issues have been formulated in order to optimize the performance of clinical predictions based on microarray data. Those issues are based on non-linear techniques and dimensionality reduction methods, taking into consideration the probability of increasing size and complexity of microarray datasets in the future. A first important conclusion from benchmarking nine microarray dataset problems is that when performing classification with least squares SVM (without dimensionality reduction), using an RBF kernel can be applied without risking overfitting on all tried datasets. The results obtained with an RBF kernel are never worse and sometimes even better than when using a linear kernel. A second conclusion is that using LS-SVM without regularization (without dimensionality reduction) ends up in very bad results, which stresses the importance of applying regularization even in the linear case. A final important conclusion is that when performing kernel PCA before classification, using an RBF kernel for kernel PCA tends to lead to overfitting, especially when using supervised feature selection. It has been observed that an optimal selection of a large number of features is often an indication for overfitting. Kernel PCA with linear kernel gives better results.

ACKNOWLEDGEMENTS

Research supported by 1. Research Council KUL: GOA-Mefisto 666, IDO (IOTA Oncology, Genetic networks), several PhD/postdoc & fellow grants; 2. Flemish Government: FWO: PhD/postdoc grants, projects G.0115.01 (microarrays/oncology), G.0240.99 (multilinear algebra), G.0407.02 (support vector machines), G.0413.03 (inference in bioi), G.0388.03 (microarrays for clinical use), G.0229.03 (ontologies in bioi), research communities (ICCoS, ANMMM); AWI: Bil. Int. Collaboration Hungary/ Poland; IWT: PhD Grants, STWW-Genprom (gene promotor prediction), GBOU-McKnow (Knowledge management algorithms), GBOU-SQUAD (quorum sensing), GBOU-ANA (biosensors); 3. Belgian Federal Government: DWTC (IUAP IV-02 (1996-2001) and IUAP V-22 (2002-2006); 4. EU: CAGE; ERNSI; 5. Contract Research/agreements: Data4s, Electrabel, Elia, LMS, IPCOS, VIB. N.P. is a research assistant of the IWT at the Katholieke Universiteit Leuven, Belgium. F.D.S is a post doctoral research assistant at the Katholieke Universiteit Leuven, Belgium. J.S. is an associate professor at the Katholieke Universiteit Leuven, Belgium. B.D.M is a full professor at the Katholieke Universiteit Leuven, Belgium.

REFERENCES

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

Bishop,C.M. (1995) *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, UK.

Brown,M.P.S., Grundy,W.N., Lin,D., Cristianini,N., Sugnet,C.W., Furey,T.S., Ares,M.Jr and Haussler,D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci.* USA, **97**, 262–267.

Cristianini,N. and Shawe-Taylor,J. (2000) *An Introduction to Support Vector Machines (and other Kernel-Based Learning Methods)*. Cambridge University Press, Cambridge.

Dawson-Saunders,B. and Trapp,R.G. (1994) *Basic & Clinical Biostatistics*. Prentice-Hall International Inc., London.

Furey,T.S., Cristianini,N., Duffy,N., Bednarski,D.W., Schummer,M. and Haussler,D. (2000) Support vector machines classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.

Golub,T.R., 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.

Gupta,H., Agrawal,A.K., Pruthi,T., Shekhar,C. and Chellappa,R. (2002) An experimental evaluation of linear and kernel-based methods for face recognition. *Workshop on the Application of Computer Vision (WACV)*, FL, USA.

Hanley,J.A. and McNeil,B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.

Hedenfalk,I., Duggan,D., Chen,Y., Radmacher,M., Bittner,M., Simon,R., Meltzer,P., Gusterson,B., Esteller,M., Raffeld,M. *et al*. (2001) Gene-expression profiles in hereditary breast cancer. *New Eng. J. Med.*, **344**, 539–548.

Iizuka,N., Oka,M., Yamada-Okabe,H., Nishida,M., Maeda,Y., Mori,N., Takao,T., Tamesa,T., Tangoku,A., Tabuchi,H. *et al*. (2003) Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. *The Lancet*, **361**, 923–929.

Mukherjee,S., Tamayo,P., Slonim,D., Verri,A., Golub,T., Mesirov,J.P. and Poggio,T. (1999) Support vector machine classification of microarray data. *A.I. Memo 1677, Massachusetts Institute of Technology*.

Nutt,C.L., Mani,D.R., Betensky,R.A., Tamayo,P., Cairncross,J.G., Ladd,C., Pohl,U., Hartmann,C., McLaughlin,M.E. *et al*. (2003) Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res.*, **63**, 1602–1607.

Pelckmans,K., Suykens,J.A.K., Van Gestel,T., De Brabanter,J., Lukas,L., Hamers,B., De Moor,B. and Vandewalle,J. (2002) LS-SVMlab: a Matlab/C Toolbox for Least Squares Support Vector Machines. *Internal Report 02-44, ESAT-SISTA, K.U.Leuven (Leuven, Belgium)*.

Schölkopf,B., Smola,A.J. and Müller,K.-R. (1998) Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, **10**, 1299–1319.

Schölkopf,B., Burges,C.J.C. and Smola,A.J. (1999) *Advances in Kernel Methods: Support Vector Learning*. MIT Press.

Schölkopf,B., Guyon,I. and Weston,J. (2001) Statistical Learning and Kernel Methods in Bioinformatics. *Proceedings NATO Advanced Studies Institute on Artificial Intelligence and Heuristics Methods for Bioinformatics*, pp. 1–21.

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

Singh,D., Febbo,P.G., Ross,K., Jackson,D.G., Manola,J., Ladd,C., Tamayo,P., Renshaw,A.A., D'Amico,A.V., Richie,J.P. *et al*. (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203–209.

Suykens,J.A.K. and Vandewalle,J. (1999) Least squares support vector machine classifiers. *Neural Process. Lett.*, **9**, 293–300.

Suykens,J.A.K., Van Gestel,T., De Brabanter,J., De Moor,B. and Vandewalle,J. (2002) *Least Squares Support Vector Machines*. World Scientific, Singapore, ISBN 981-238-151-1.

Suykens,J.A.K., Van Gestel,T., Vandewalle,J. and De Moor,B. (2003) A support vector machine formulation to PCA analysis and its Kernel version. *IEEE T. Neural Network.*, **14**, 447–450.

Van Gestel,T., Suykens,J.A.K., Lanckriet,G., Lambrechts,A., De Moor,B., Vandewalle, J. (2002) Bayesian framework for least squares support vector machine classifiers, Gaussian processes and kernel Fisher discriminant analysis. *Neural Comput.*, **15**, 1115–1148.

Van Gestel,T., Suykens, J.A.K., Baesens,B., Viaene,S., Vanthienen,J., Dedene,G. De Moor,B. and Vandewalle, J. (2004) Benchmarking least squares support vector machine classifiers. *Mach. Learn.*, **54**, 5–32.

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. *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Vapnik,V.N. (1998) *Statistical Learning Theory*. John Wiley & Sons, New York.