BIOINFORMATICS

Vol. 19 no. 0 2003, pages 1–8
DOI: 10.1093/bioinformatics/btg062

Effective dimension reduction methods for tumor classification using gene expression data

A. Antoniadis*, S. Lambert-Lacroix and F. Leblanc

Laboratoire IMAG-LMC, University Joseph Fourier, BP 53, 38041 Grenoble Cedex 9, France

Received on April 21, 2002; revised on October 8, 2002; accepted on November 10, 2002

ABSTRACT

Motivation: One particular application of microarray data, is to uncover the molecular variation among cancers. One feature of microarray studies is the fact that the number n of samples collected is relatively small compared to the number p of genes per sample which are usually in the thousands. In statistical terms this very large number of predictors compared to a small number of samples or observations makes the classification problem difficult. An efficient way to solve this problem is by using dimension reduction statistical techniques in conjunction with nonparametric discriminant procedures.

Results: We view the classification problem as a regression problem with few observations and many predictor variables. We use an adaptive dimension reduction method for generalized semi-parametric regression models that allows us to solve the ‘curse of dimensionality problem’ arising in the context of expression data. The predictive performance of the resulting classification rule is illustrated on two well know data sets in the microarray literature: the leukemia data that is known to contain classes that are easy ‘separable’ and the colon data set.

Availability: Software that implements the procedures on which this paper focus are freely available at http://www-lmc.imag.fr/SMS/software/microarrays/.

Contact: Anestis.Antoniadis@imag.fr

INTRODUCTION

The microarray technology makes it now possible to rapidly measure, through the process of hybridization, the levels of virtually all the genes expressed in a biological sample. The gene expression patterns in microarray data have already provided some valuable insights in a variety of problems, and it is expected that knowledge gleaned from microarray data will contribute significantly to advances in fundamental questions in biology as well as in clinical medicine.

One particular application of microarray data, is to uncover the molecular variation among cancers. Classification of different cell types, predominantly cancer types, using microarray gene expression data has been considered by Golub *et al.* (1999) for classification of acute leukemia, Alon *et al.* (1999) for cluster analysis of tumor and normal colon tissues, and Alizadeh *et al.* (2000), for diffuse large B-cell lymphoma (DLBCL), to cite only a few. The methods used in most of the above papers range from discriminant analysis over Bayesian approaches to other analysis techniques from machine learning such as boosting, bagging and support vector machines. A comprehensive comparative study of several discrimination methods and machine learning methods in the context of cancer classification based on filtered sets of genes can be found in Dudoit *et al.* (2002).

One feature of microarray studies is the fact that the number n of samples collected is relatively small compared to the number p of genes per sample which are usually in the thousands. In statistical terms the very large number of predictors or variables (genes) compared to a small number of samples or observations (microarrays) make most of classical ‘class prediction’ methods difficult to employ, unless a preliminary variable selection step is performed. For example, the pooled within-class sample covariance matrix required to form Fisher’s linear discriminant function is singular if $n < p + 2$. Even if all the genes can be used as, say, with a Euclidean-based rule or a support vector machine, the use of all the genes allows presence of the noise associated with genes of little or no discriminatory power, and this inhibits and degrades the performance of the classification rule in its application to unclassified tumors. That is, although the apparent error rate of the classification rule (the proportion of the training tissues misallocated by the rule) will decrease as it is formed from more and more genes, its error rate in classifying tissues outside of the training set eventually will increase.

In most of the previous studies mentioned, the authors have used univariate methods for reducing the number of genes to be considered before using appropriate classification procedures. An alternative approach to handle the ‘small n, large p’ problem is the one used by West *et*

*To whom correspondence should be addressed.

al. (2001) based on a Bayesian probit binary regression model for a data for which there is a binary clinical outcome response variable and the predictors consist of the gene expression levels. West *et al.* (2001) used techniques based on the singular value decomposition (SVD) of the $p \times n$ matrix whose columns are the gene expression profiles of the n different tumors. Nguyen and Rocke (2002) and Gosh (2002) proposed using the method of partial least squares (PLS) for dimension reduction, as a preliminary step to classification using linear logistic discrimination (LD), linear (LDA) or quadratic discriminant analysis (QDA).

The purpose of the SVD used by West *et al.* (2001) is to produce orthogonal tumor descriptors that reduce the high dimensional data to only a few gene components (super genes) which explain as much of the observed total gene expression variation as possible. However, this is achieved without regards to the response variation and may be inefficient. One simple explanation is that this way of reducing the regressor dimensionality is totally independent of the output variable. Thus any two different data sets would always reduce to the same linear combinations, as long as the input variables have the same distributions. This is so, even if the relationship between the predictors and the response is not the same for the two data sets. To address the dimension reduction issue, one must not treat the predictors separately from the response. This is the spirit of the methods developed by Nguyen and Rocke (2002) and Gosh (2002), where the PLS components are chosen so that the sample covariance between the response and a linear combination of the p predictors (genes) is maximum. One may argue that the latter criterion for PLS is more sensible since there is no a priori reason why constructed components having large predictor variation (gene expression variation) should be strongly related to the response variable. However, PLS is really designed to handle continuous responses and especially for models that do not really suffer from conditional heteroscedasticity as it is the case for binary or multinomial data. The only attempt, to our knowledge, to extend PLS procedures to categorical response variables is a Proceedings paper by Gauchi and Tenehaus (1994). Their approach consists in transforming first the categorical variable into a continuous one via multiple correspondence analysis and to use then standard PLS for continuous response. The question is then in what sense, can the PLS reduction from the original p predictors to the first few components can be effective? Are there any other linear combinations more useful than the PLS components in providing useful information about group separation?

In this article, we address these issues by formulating the classification problems via the adaptive effective dimension reduction approach (MAVE) of Xia *et al.* (2002) for general regression problems, extending Li's (1991) sliced inverse regression (SIR) methods. We use the MAVE method as proposed in the second half of the paper by Xia *et al.* (2002) and which is an extension of the least-squares MAVE procedure to the class of generalized nonlinear models which includes qualitative response models and in particular binary, multinomial, ordered response, and other discrete choice models. Then, to capture nonlinear structures between the expected binary responses and the canonical variates so obtained without the knowledge about the functional relationship in advance, we use parametric and nonparametric logistic discrimination rules following the dimension reduction. Our procedure can be viewed as a two-stage procedure. The first stage (feature selection) is to find the canonical variates for reducing the predictor dimension from p to some integer much smaller than the number of observations; the second stage is to split the canonical space into 2 regions for class-membership prediction via appropriate parametric or nonparametric discriminant rules. The combination of such discriminant rules and the dimension reduction achieved by our method yields good prediction results.

This paper is organized as follows. In the Methods section we describe the dimension reduction methods of SIR and MAVE for binary data, the classification methods of parametric and nonparametric logistic discrimination and a preliminary gene selection strategy based on a unsupervised fold change followed by Wilcoxon like score based gene selection procedure as defined in Park *et al.* (2001). We then apply our algorithms to two publicly available microarray data sets which have been considered before by several authors. Misclassification rates for our classifiers are estimated using leave-one-out cross-validation on the training set.. The results from applying our method and their competitive performance comparison with other methods are given in the Results section. We end with some discussion and concluding remarks. An appendix is included presenting a brief discussion of the computation algorithms involved and their implementation.

METHODS

Discriminant analysis aims at the classification (supervised learning) of an object into one of K given classes based on information from a set of p predictor variables observed on n samples. When focusing on the classification of tumors using gene expression data, traditional methods such that LDA or QDA or LD do not work since $p + 2 > n$. Thus, methods able to cope with the high dimensionality of the data are needed. In this section, after briefly recalling the theoretical foundations that support traditional classification methods, we set up the classification problem under the effective dimension reduction binary regression model. Through the connection between various approaches for dimension reduction in generalized regression models our methodology leads to several ways of generalizing LD for better exploration and exploitation of nonlinear data patterns in microarray data.

Classifiers

In this subsection we briefly describe some supervised classification methods that groups classifiers into two main categories according to their underlying mathematical principles. A more complete overview of such methods is given in Dudoit *et al.* (2002). To fix the notation, gene expression data on p genes for n mRNA samples will be summarized by an $p \times n$ matrix $X = (x_{ij})$, where x_{ij} denotes the expression level of gene (variable) i in mRNA sample (observation) j. The expression levels might be either absolute (e.g. oligonucleotide arrays that were used to produce the leukemia dataset) or relative with respect to the expression levels of a suitably defined common reference sample (e.g. cDNA microarrays). In supervised classification, each mRNA sample is thought to originate from a specific class $k \in \{1, \ldots, c\}$, where the number of possible classes c is known and fixed. The data for each observation consist of a gene expression profile (pattern) $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{pj})^T$ and a class label y_j, i.e. under classical regression terminology, of predictor vector valued variables $\boldsymbol{x}_j$ and response y_j. Hereafter, we shall assume that the pairs $(\boldsymbol{x}_j, y_j)$, $j = 1, \ldots, n$ are independent and identically distributed realizations of a random vector $(\boldsymbol{X}, Y)$, although taking sometimes advantage of context dependent information might be beneficial. We let n_k denote the number of observations belonging to class k.

A classifier can be regarded as a function $g : \mathbb{R}^p \to \{1, \ldots, c\}$ that predicts the unknown class label of new tissue sample $\boldsymbol{x}$ by $g(\boldsymbol{x})$. The *a priori* probability of class k is $P_k = \mathbb{P}(Y = k)$ and the conditional probability density of $\boldsymbol{X}$ for class k is denoted by f_k. The pooled data with all the classes combined has then the density function $f = \sum_{k=1}^{c} P_k f_k$. Sometimes the class-conditional densities f_k are supported by disjoint regions in the space $\mathbb{R}^p$ and it is then possible to construct an optimal classifier with zero classification error $\mathbb{P}(g(\boldsymbol{X}) \neq Y)$. More often however the individual classes overlap and the smallest achievable misclassification error is positive. If the probabilities P_k and the class-conditional densities f_k were known, the classifier that minimizes the misclassification risk is called the *Bayes classifier* and is defined by

$$g_{\text{Bayes}}(\boldsymbol{x}) = \text{argmax}_{k \in \{1,\ldots,c\}} P_k f_k(\boldsymbol{x}). \tag{1}$$

An alternative way is to consider the *a posteriori* probability $\mathbb{P}(Y = k|\boldsymbol{X})$ and use the rule

$$g_{\text{Bayes}}(\boldsymbol{X}) = \text{argmax}_{k \in \{1,\ldots,c\}} \mathbb{P}(Y = k|\boldsymbol{X}). \tag{2}$$

Both these rules are equivalent.

In practice, the class-conditional densities or the conditional probabilities $\mathbb{P}(Y = k|\boldsymbol{X})$ in the above classification rules are built from past experience, i.e. are estimated from observations which are known to belong to certain classes. Such observations comprise the learning set (LS). Predictors may then be applied to a test set (TS) to predict for each observation $\boldsymbol{x}_j$ in the test set its class y_j. In the event that the y_j are known, the predicted and true classes may be compared to estimate the error rate of the predictor. To estimate the Bayes rule two distinct approaches emerge. The use of rule (1) requires explicit estimation of the class-conditional densities f_k. For rule (2), regression techniques may be used to estimate the posterior probabilities $\mathbb{P}(Y = k|\boldsymbol{X})$ without considering the class-conditional densities separately. Both these approaches are feasible, under parametric or nonparametric assumptions when $p << n$, but the case $p >> n$ is a formidable challenge. A promising way to meet it is by using nonparametric classification techniques in conjunction with optimal dimension reduction techniques. In what follows, we will focus first on binary problems with response $Y \in \{0, 1\}$.

Dimension reduction through regression

The final goal of a regression analysis is to understand how the conditional distribution of a univariate response Y given a vector $\boldsymbol{X}$ of p predictors depends on the value of $\boldsymbol{X}$. We have already seen in the previous subsection that if the conditional distribution of $Y|\boldsymbol{X}$ was completely known for each value of $\boldsymbol{X}$ then the classification problem would be definitely solved. However, under our setting, the study of $Y|\boldsymbol{X}$ is problematic since the number of available observations is small and the dimension of $\boldsymbol{X}$ is extremely large. Such difficulties can be addressed by placing restrictions on $Y|\boldsymbol{X}$ and by limiting the regression objective to specific characteristics of $Y|\boldsymbol{X}$. Such a task may be achieved by dimension reduction procedures *without loss of information*, and without requiring a specific model for $Y|\boldsymbol{X}$. These procedures have the potential to point to useful classifiers.

In binary regression, without prior knowledge about the relationship between Y and $\boldsymbol{X}$, the regression function

$$r(\boldsymbol{x}) = \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x})$$

or the logit function

$$s(\boldsymbol{x}) = \log\left(\frac{\mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x})}{\mathbb{P}(Y = 0|\boldsymbol{X} = \boldsymbol{x})}\right)$$

are often modelled in a flexible nonparametric fashion. When the dimension of $\boldsymbol{X}$ is high, recent efforts have been expended in finding the relationship between Y and $\boldsymbol{X}$ efficiently. The final goal is to approximate $r(\boldsymbol{x})$ or $s(\boldsymbol{x})$ by a function having simplifying structure which makes estimation and interpretation possible even for moderate sample sizes. There are essentially two approaches: the first is largely concerned with function approximation and the second with dimension reduction. Examples of the former are the generalized additive model approach of Hastle and Tibshirani (1986) and the projection pursuit regression proposed by Friedman and Stuotzle (1981); both assume that the function to be estimated is a sum of univariate smooth functions. Examples of the latter are the dimension reduction of Li (1991) and the regression graphics of Cook (1994) and more lately the adaptive approach based on semiparametric models of Xia *et al.* (2002). One goal of this article is to implement the procedures of Xia *et al.* (2002) to our classification set-up.

A regression-type model for dimension reduction can be written as

$$Y = g(B_0^T \boldsymbol{X}) + \epsilon \tag{3}$$

where g is an unknown smooth link function, $B_0 = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_D)$ is is a $p \times D$ orthogonal matrix ($B_0^T B_0 = I_D$) with $D < p$ and $\mathbb{E}(\epsilon|\boldsymbol{X}) = 0$ almost surely. The last condition allows ϵ to be dependent on $\boldsymbol{X}$ and covers, in particular, the binary regression case. In the terminology of Cook and Weisberg (1999) the above model implies that the distribution of $Y|\boldsymbol{X}$ is the same as that of $Y|B_0^T\boldsymbol{X}$ and therefore the p-dimensional predictor $\boldsymbol{X}$ can be replaced by the D-dimensional predictor $B_0^T\boldsymbol{X}$ without loss of regression information, and thus represents a potentially useful reduction in the dimension of the predictor vector. Of course, in our case this is a simplifying structural assumption which serves us to identify a small number of linear combinations of a subset of genes that can be used to predict cancer tumor genotype. The space spanned by the columns of the matrix B_0, can be uniquely defined under some mild conditions and is called the *effective dimension reduction* (EDR) space. As in Xia *et al.* (2002) we shall refer to the column vectors of B_0 as EDR directions, which are unique up to orthogonal transformations. The estimation of the EDR space includes the estimation of the directions, namely B_0, and the corresponding dimension of the EDR space.

A brief review of several specific semiparametric methods to estimate B_0 are given in Xia *et al.* (2002). When the predictors are elliptically distributed and a constant variance among classes holds, the sliced inverse regression (SIR) method proposed by Li (1991) and the sliced average variance estimation (SAVE) proposed by Cook (1994) are perhaps, up to now, the most powerful methods for searching for EDR directions and dimension reduction. However, it can be shown (see Kent, 1991; Li, 2000, ch. 14 and Cook and Lee, 1999) that for binary data the SIR method is equivalent to linear discriminant analysis, the only difference being the way the discrimination directions are scaled, and that it is fair to think of SAVE as the quadratic discriminant analysis subspace (QDA) in the same way that SIR corresponds to the LDA subspace.

In this paper, we have used the new method proposed by Xia *et al.* (2002) to estimate the EDR directions. It is called the (conditional) minimum average variance estimation (MAVE) method. It is easy to implement and needs no strong assumptions on the probabilistic structure of $\boldsymbol{X}$. It is out of the scope of the present paper to describe in details the computations underlying the derivation of the MAVE method. These are given in full extend in the above cited paper. Let us only mention that the MAVE method may be seen as a combination of nonparametric function estimation by local polynomials and direction estimation, which is executed simultaneously with respect to the directions and the nonparametric link function. Moreover, under the restriction that, under model (3), $\boldsymbol{X}$ has a density with compact support, the authors of the above cited paper show, by extending the cross-validation method of Cheng and Tong (1992), that the dimension of the EDR space can be consistently estimated.

Classification: nonparametric logistic regression and local density estimation

After dimension reduction by MAVE, the high dimension of p is now reduced to a lower dimension of D super-gene components. Once the D components are constructed we consider prediction of the response classes. Since the reduced (gene) dimension is now low ($D << n$), conventional classification methods may be used.

Let $\boldsymbol{z}$ be an observed column vector of D super-genes predictor values. In conventional linear logistic regression, the conditional class probability, $\mathbb{P}(Y|\boldsymbol{Z} = \boldsymbol{z})$ is modeled using the logistic functional form

$$\pi(\boldsymbol{z}) = \frac{\exp(\xi_0 + \boldsymbol{z}^T\boldsymbol{\xi})}{1 + \exp(\xi_0 + \boldsymbol{z}^T\boldsymbol{\xi})}, \tag{4}$$

where the constant ξ_0 and the D-dimensional parameter vector $\boldsymbol{\xi}$ are estimated by maximum likelihood using Fisher scoring. The predicted response probabilities are estimated by replacing ξ_0 and $\boldsymbol{\xi}$ with their maximum likelihood estimators $\hat{\xi}_0$ and $\hat{\boldsymbol{\xi}}$ and the classifier predicts the value 1 for a new sample if its estimated conditional probability is larger than 0.5. This classification procedure is called logistic discrimination (LD). The advantage of the linear-logistic discrimination lies not only in its computational convenience, but more importantly in the ease of interpretation of the model parameters, and our ability to make inference about them. However, while often a linear-logistic model fits the data reasonably well, sometimes there might be some curvature in the logit that is not captured by it. A simple nonparametric alternative to the fully parametric model (4) which allows for such curvature, but yet retains the ease of interpretation of parameters such as $\boldsymbol{\xi}$ is the model

$$\pi(\mathbf{z}) = \frac{\exp(\eta(\mathbf{z}^T\boldsymbol{\xi}))}{1+\exp(\eta(\mathbf{z}^T\boldsymbol{\xi}))}, \tag{5}$$

for some completely unknown smooth function η. Such models fall under the class of generalized linear single-index models and efficient methods that deal with fitting and making inference about such models have been developed by Carroil *et al.* (1997) (see also Hastle and Tibshirani (1986) for additive generalized binomial models).

As we have already noticed, other classification methods are obtained using the density estimation approach. For these one needs estimates for the prior probabilities P_k as well as the conditional densities f_k in the Bayes classifier. The prior probabilities may either be known or they can be estimated from the relative frequencies of the classes among the training data. The difficult part is to estimate the class-conditional densities. A classical approach is to model the class-conditional densities as multivariate normal distributions and this leads to quadratic discriminant analysis (QDA). In case of equal covariance matrices one obtains linear discriminant analysis (LDA).

Modeling the class-conditional densities as multivariate normals is an example of parametric density estimation, where the densities are assumed to belong to a family of functions described by a finite set of parameters. In nonparametric density estimation no such fixed family of possible densities is assumed. Kernel or local polynomial nonparametric estimates are then used but unless the dimension of the support of the densities is small these methods suffer from what is called the curse of dimensionality. Such nonparametric methods are perfectly feasible, if one replaces the full gene profiles by their projections onto the EDR space. However, in the study that follows, learning and training test sets are not sufficiently large to provide adequate density estimation for density based nonparametric discrimination.

Preliminary Gene Selection

The intrinsic problem with classification from microarray data is that sample size is much smaller than the number p of genes. Theoretically MAVE capitalizes on the correlations among the genes and with the class labels to identify a small number of linear combinations of a subset of genes that can be used to predict cancer tumor genotype via parametric or nonparametric logistic classification. As such it is a method that can handle a large number of genes. However, as many other multivariate methods it is challenged by severe memory requirements, large computational time, singular empirical covariance matrices and the danger of over-fitting. The traditional attempt to overcome these problems consists in data compression methods. Therefore, we have used, as suggested by a referee and as it is done in several other papers discussing class prediction with gene expression data two preliminary steps for compression: unsupervised fold change (as it is done in particular in Dudoit *et al.* (2002) followed by the Park *et al.* (2001) Wilcoxon score based gene selection procedure.

RESULTS

We demonstrate the usefulness of the proposed methodology described above to two well known data sets: the leukemia data first analyzed in Golub *et al.* (1999) and the colon data analyzed initially by Alon *et al.* (1999). Both data sets consist of absolute measurements from Affymetrix high-density oligonucleotide arrays: the first contains $n = 72$ tissue samples on $p = 7129$ genes (47 cases of acute lymphoblastic leukemia (ALL) and 25 cases of acute myeloid leukemia (AML)) and the second $n = 62$ tissue samples on $p = 2000$ human gene expressions (40 tumors and 22 normal tissues). We have used the MATLAB software environment for preprocessing the data and to implement our proposed classification methodology. The suite of MATLAB functions implementing our procedure is freely available at the URL http://www-lmc.imag.fr/SMS/Software/mi-croarrays/.

As it is common we assess the performance of the classification rules for a selected subset of genes by their errors on the test set and also by their leave-one-out cross-validated errors. But, as pointed out by the referees, if these errors are calculated within the gene preliminary selection process, there is a selection bias in them when they are used as an estimate of the prediction error. Therefore, to fairly evaluate and compare the test error or the leave-one-out cross-validated error of the classification rules that follow, we perform gene selection in training the rules at each stage of the cross-validation process using only the training sample at hand, following a methodology similar to the one that has been considered by Nguyen and Rocke (2002). Of course, when leave-one-out cross-validation is used there is no guarantee that the same subset of genes will be obtained as during the original training of the rule (on all the training observations). Indeed, with the huge number of genes available, it generally will yield a subset of genes that has at most only a few genes in common with the subset selected during the original training of the rule. The leave-one-out CV error is nearly unbiased, but it can be highly variable.

We first describe the results on the leukemia data set (available at the URL http://www.genome.wi.mit.edu/MPR). We followed exactly the protocol in Dudoit *et al.* (2002) to pre-process the data by thresholding, filtering, a base 10 logarithmic transformation and standardization, so that the final data is summarized by a 3571 × 72 matrix $X = (x_{ij})$, where x_{ij} denotes the base 10 logarithm of the expression level for gene i in mRNA sample j. In this study, the data are already divided into a learning set of 38 mRNA samples and a test set of 34 mRNA samples. The observations in the two sets came from different labs and were collected at different times. The test set comprises a broader range of samples, including samples from peripheral blood as well as bone marrow, from childhood AML patients, and from laboratories that used different sample preparation protocols.

When training the rule and for the pre-selection of genes, we first reduced the set of available genes to the top $p^* = 50$, 100 and 200 genes as ranked in terms of Wilcoxon score based Park *et al.* (2001) procedure. Another variable selection criterion for gene selection in two-class prediction is the one based on a BSS/WSS criterion and used by Dudoit *et al.* (2002) which we also tried, and which produces similar effects on performance. To compare our results we have also applied the two discriminant analysis procedures DLDA et DQDA described in Dudoit *et al.* (2002), which implement a simple Gaussian maximum likelihood discriminant rule, for diagonal class covariance matrices, assumed to be constant across classes for DLDA and varying across classes for the DQDA rule. The reason for this comparison is that such diagonal procedures are not subject to the curse of dimensionality and moreover DLDA was pointed by Dudoit *et al.* (2002) to have remarkably low error rates in their study of a subset of the same dataset. In order to mimic the prediction of a new observation from a training sample, we used a cross-validation estimate of the misclassification rate for every method, where each observation of the training set is successively omitted from the data, and then classified with gene selection and classification based on the remaining observations. The results are given in Table 1.

The estimated EDR dimension D was most of the time equal to 1 ($D = 2$ for samples 6, 17 and 22, when $p^* = 50$). As one can see, our method, using a parametric (LD) or nonparametric logistic discrimination (NPLD) predicts well the ALL/AML classes for the 38 training samples using leave-one-out bias gene selection corrected cross-validation. One exception is the ALL sample 17, which is missclassified when $p^* = 50$. Given that the estimated EDR dimension for most cross-validation runs was $D = 1$, there was no gain in using nonparametric logistic discrimination in this case. However this was not true for the test set. For the same cross-validation runs the DLDA procedure predicted all classes correctly, while the DQDA procedure had one or two misclassifications in the bias selection corrected cross-validated learning set.

Table 1. Classification rates by the four methods for the leukemia data set with 38 training samples (27 ALL, 11 AML) and 34 test samples (20 ALL, 14 AML). Given are the number of correct classification out of 38 and 34 for the training and test samples respectively.

p^*	MAVE-LD	DLDA	DQDA	MAVE-NPLD
Training Data (Leave-out-one CV)				
50	37	38	37	37
100	38	38	37	38
200	38	38	36	38
Test Data (Out-of-sample)				
50	33	33	33	33
100	33	33	33	33
200	32	33	32	32

Table 2. Error rates of the classification rules with the Park *et al.* (2001) nonparametric scoring selection algorithm, averaged over 50 random splits of the 72 leukemia tissue samples into training and test subsets of 38 and 34 samples, respectively.

p^*	MAVE-LD	DLDA	DQDA
50	0.0571 (0.048)	0.0406 (0.028)	0.0321 (0.029)
100	0.0335 (0.028)	0.0347 (0.028)	0.0312 (0.025)
200	0.0241 (0.025)	0.0288 (0.029)	0.0224 (0.023)

Prediction on the test samples using parametric or nonparametric logistic discrimination on the $D = 1$ training MAVE components resulted in one misclassification for $p^* = 50$ or 100, and 2 for $p^* = 200$, while both DLDA and DQDA procedures resulted in a 1 or 2 over 34 misclassified samples. These results can also be directly compared to those obtained in the study of Golub *et al.* (1999), where 29 observations were correctly classified by their voting scheme and the results by Furey *et al.* (2000) working with support vector machines, reporting results ranging between 30 and 32 correct classifications in the test samples.

To further investigate the bias and variability of the prediction rules, we split 50 times the set of 72 tissues into a training set of 38 tissues (25 ALL and 13 AML) and a test set of 34 tissues (22 ALL and 12 AML) by sampling without replacement from the 47 ALL and 25 AML samples separately. For each of the 50 splits, the training set is used to carry out gene selection and an unbiased error rate estimate is given by the test error, equal to the proportion of tissues in the test set mis-allocated by the rule. The average values of the error rate estimates and their standard deviations (in parentheses) are reported in Table 2. It can be seen from Table 2 that the prediction error improves with the number of retained genes and that the MAVE (combined with logistic discrimination) based rates range between DLDA and DQDA. However, note that DLDA appears to be worse than DQDA on the average of the 50 splits.

To further illustrate the method we have also applied the same procedures to the colon data set (see Table 3).

Table 3. Classification rates by the three methods for the colon data set. Given are the number of correct classification out of 62 with leave-out-one cross-validation.

p^*	MAVE-LD	DLDA	DQDA
Data (Leave-out-one CV)			
50	52	51	51
100	52	51	51
200	52	51	51

Table 4. Error rates of the classification rules with the Park *et al.* (2001) nonparametric scoring gene selection algorithm, averaged over 50 random splits of the 62 colon tissue samples into training and test subsets of 31 samples each.

p^*	MAVE-LD	DLDA	DQDA
50	0.1392 (0.056)	0.1432 (0.047)	0.1400 (0.053)
100	0.1315 (0.046)	0.1406 (0.043)	0.1355 (0.051)
200	0.1336 (0.045)	0.1561 (0.045)	0.1565 (0.048)

From the table above one can see that the classes in the colon data set are less well separated and that all methods perform equivalently. Note that the rates supplied here seem to be much worse than those given in Table 8 of Nguyen and Rocke (2002), but we suspect that their table displays the rates obtained when applying their classification rule to the training sample without cross-validation correction. Indeed, using the training sample as a test sample without cross-validation we obtain at most 3 missclassified samples, whatever the value of p^* is.

For these data and in order to investigate the bias and variability of the prediction rules, we split it into a training and a test set each of size 31 by sampling without replacement from the 40 tumor and 22 normal tissues separately such that each set contained 20 tumor and 11 normal tissues. Here again the training set was used to carry out gene selection and form the error rate estimates for three sizes of selected subset of genes. We calculated these quantities for 50 such splits of the colon data into training and test sets. The results are reported in Table 4. It can be seen that MAVE combined with LD is less biased that the two other methods but the mean squared error of the three methods are equivalent.

CONCLUSIONS

We have proposed a statistical dimension reduction approach for the classification of tumors based on microarray gene expression data. Our method is designed to address the curse of dimensionality to overcome the problem of a high dimensional gene expression space so common in such type of problems. To apply the methodology we viewed the classification problem as a regression problem, with group membership being the response variable, and used adapted parametric or nonparametric regression methods to solve the problem. The results on two real data sets show that such an approach is successful. While we have not illustrated the methodology for multiclass problems we believe that our approach can be also help by reducing the multiclass problem to multiple binary problems that are solved separately as it is usually done in the machine learning literature. The nonparametric logistic classifier after dimension reduction is suitable for gene expression arrays of any size, but we believe that, with the number of experiments growing rapidly, for large sample arrays nonparametric density based discrimination after dimension reduction can be proven as powerful.

ACKNOWLEDGEMENTS

This work was supported by the research project GEN-IMAG. The authors would also like to thank Yingcun Xia, Howell Tong, W. K. Li and Li-Xing Zhu for providing a preprint of their paper prior to publication and also a set of Matlab functions implementing their procedure in additive noise models. They would also like to thank Avner Bar-Hen for his comments on a preliminary version of this paper and the three referees and the associated editor for their constructive comments and criticisms which have substantially improved this article.

REFERENCES

Alizadeh,A., Eisen,M., Davis,R., Ma,C., Lossos,I., Rosenwald,A., Broldrick,J., Sabet,H., Tran,T., Yu,X. *et al.* (2000) Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Alon,U., Barkai,N., Notterman,D., 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. *PNAS*, **96**, 6745–6750.

Carroll,R., Fan,J., Gijbels,I. and Wand,M.P. (1997) Generalized partially linear single-index models. *J. Amer. Statist. Assoc.*, **92**, 477–489.

Cheng,B. and Tong,H. (1992) On consistent nonparametric order determination and chaos (with discussion). *J. R. Statist. Soc. B*, **54**, 427–449.

Cook,R.D. (1994) On the interpretation of regression plots. *J. Amer. Statist. Assoc.*, **94**, 177–189.

Cook,R.D. and Lee,H. (1999) Dimension reduction in regressions with a binary response. *J. Amer. Statist. Assoc.*, **94**, 1187–1200.

Cook,R.D. and Weisberg,S. (1999) *pplied Regression Including Computing and Graphics*. Wiley, New York.

Dudoit,S., Fridlyand,J. and Speed,T. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Statist. Assoc.*, **97**, 77–87.

Friedman,J.H. and Stuotzle,W. (1981) Projection pursuit regression. *J. Amer. Statist. Assoc.*, **76**, 817–823.

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

Gauchi,J.-P. and Tenehaus,M. (1994) Régression PLS qualitative of application à un problème de formulation d'une huile silicone fonctionnalisée. *26 meeting of the French Statistical Society*, ASU, 24-27 May 1994, Neuchâiel, CH.

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

Gosh,D. (2002) Singular value decomposition regression modelling for classification of tumors from microarray experiments. In *Proceedings of the Pacific Symposium on Biocomputing*. In Press.

Hastle,T.J. and Tibshirani,R. (1986) Generalized additive models (with discussion). *Statist. Sci.*, **1**, 336–337.

Kent,J.T. (1991) Discussion of li. *J. Amer. Statist. Assoc.*, **86**, 336–337.

Li,K.C (1991) Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.*, **86**, 316–342.

Li,K.C. (2000) *High dimensional data analysis via the sir/phd approach*. Unpublished manuscript dated April 2000.

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

Park,P., Pagano,M. and Bonetti,M. (2001) A nonparametric scoring algorithm for identifying informative genes from microarray data. In *Pacific Symposium on Biocomputing*. pp. 52–63.

West,M., Blanchette,C., Dressman,H., Huang,F., Ishida,S., Spang,R., Zuzan,H., Olason,J., Marks,I. and Nevins,J. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *PNAS*, **98**, 11462–11467.

Xia,Y., Tong,H., Li,W.K. and X,Z.L. (2002) An adaptive estimation of dimension reduction space. *J. R. Statist. Soc. B.*, **64**, 363–410.

APPENDIX

We briefly describe here the main steps of the MAVE algorithm considered in Xia *et al.* (2002) which is devoted to the estimation of the matrix B in $\mathbb{E}(Y|\boldsymbol{X}) = g(B^T\boldsymbol{X})$ allowing g to be an unknown smooth function. The estimated B is a solution to the problem

$$\min_B \mathbb{E}\{Y - \mathbb{E}(Y|B^T\boldsymbol{X})\}^2 = \mathbb{E}(\sigma_B^2(B^T\boldsymbol{X})),$$

subject to $B^TB = I$. To minimize the above expression one has first to estimate the conditional variance $\sigma_B^2(B^T\boldsymbol{X}) = \mathbb{E}[\{Y - \mathbb{E}(Y|B^T\boldsymbol{X})\}^2|B^T\boldsymbol{X}]$. Let $g_B(\boldsymbol{v}) = \mathbb{E}(Y|B^T\boldsymbol{X} = \boldsymbol{v})$. Given a sample $\{\boldsymbol{X}_i, Y_i\}$ a local linear fit is applied to estimate $g_B(\cdot)$ and the EDR directions are estimated by solving the minimization problem

$$\min_{B,a_j,\boldsymbol{b}_j} \left(\sum_{j=1}^{n} \sum_{i=1}^{n} (Y_i - [a_j + \boldsymbol{b}_j^T B^T (X_i - X_j)])^2 w_{ij} \right), \tag{6}$$

where $w_{ij} = K_h\{\boldsymbol{B}^T(X_i - X_j)\}/\sum_{\ell=1}^n K_h\{\boldsymbol{B}^T(X_\ell - X_j)\}$ are multidimensional kernel weights. We start with the identity matrix as an initial estimator of B to be used in the kernel weights. Then iteratively, we use the multidimensional kernel weights to obtain an estimator $\hat{\boldsymbol{B}}$ by minimizing problem (6) and refine the kernel weights with the updated value of B and iterate until convergence. The choices of the bandwidth h and the EDR dimension d are implemented through a cross-validation technique.

To be balanced at final stage