DE GRUYTER OPEN

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 43 (56) 2015

DOI: 10.1515/slgr-2015-0043

Visualization and Comparison of Single and Combined Parametric and Nonparametric Discriminant Methods for Leukemia Type Recognition Based on Gene Expression

Małgorzata M. Ćwiklińska-Jurkowska[1]

[1] Department of Theoretical Foundations of Biomedical Sciences and Medical Computer Science, Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University, Poland

Abstract. A gene expression data set, containing 3051 genes and 38 tumor mRNA training samples, from a leukemia microarray study, was used for differentiation between ALL and AML groups of leukemia. In this paper, single and combined discriminant methods were applied on the basis of the selected few most discriminative variables according to Wilks' lambda or the leave-one-out error of first nearest neighbor classifier. For the linear, quadratic, regularized, uncorrelated discrimination, kernel, nearest neighbor and naive Bayesian classifiers, two-dimensional graphs of the boundaries and discriminant functions for diagnostics are presented. Cross-validation and leave-one-out errors were used as measures of classifier performance to support diagnosis coming from this genomic data set. A small number of best discriminating genes, from two to ten, was sufficient to build discriminant methods of good performance. Especially useful were nearest neighbor methods. The results presented herein were comparable with outcomes obtained by other authors for larger numbers of applied genes. The linear, quadratic, uncorrelated Bayesian and regularized discrimination methods were subjected to bagging or boosting in order to assess the accuracy of the fusion. A conclusion drawn from the analysis was that resampling ensembles were not beneficial for two-dimensional discrimination.

Introduction

Genomic research is the foundation upon which the rules of personalized medicine are developed. An important issue to consider related to this topic is how various diseases modify gene expression. The microarray technique presents an opportunity for a more efficient approach to classification, using simultaneous observation of gene expression via DNA microarrays. Many studies on microarrays concern this subject, e.g. classical works under Golub et al. (1999) and Marchiori et al. (2005). Tumors are very often a cause of death. Thus, an especially large amount of work in the analysis of microarray experiments is concerned with research on cancers

ISBN 978–83–7431–464–0 ISSN 0860-150X

(e.g. Marchiori et al., 2005; Pomeroy et al., 2002). Accurate identification of cancer type is often essential for successful treatment. For example, different types of leukemia are treated in different ways. Thus, it might be useful to use discriminant analysis to support diagnosis. In exploring this topic, a leukemia data set containing expression levels of genes is applied in the current work.

Discriminant analysis is often used to determine which variables are the best predictors of classification. Moreover, it can be applied to assign observations to categories or to groups. For a set of observations containing values of variables and classification information defining groups of elements, every discriminant method supplies a criterion to classify each case. Due to the "curse of dimensionality" in microarray analysis, most standard statistical methods might not be useful. Because of the high number of investigated genes in one microarray, the pre-selection of features for inclusion into the classification rule is essential. Often, only a few tens of genes are really active; the remaining genes are not important for improvement of the discriminant procedure. In supervised classification, the variables with the biggest discriminant power are sought out. We search for genes useful for differentiation, without a significant decrease of information coming from the data. Medical decisions may be supported according to the classification model, which is based on chosen variables. It is interesting to discover which, of several potential discriminant methods, have the lowest rates of misclassification.

Classical techniques (supplying parametric discriminant functions) assume joint normality of predictive variables. However, in many cases this assumption, or assumption of equal covariance matrices for quadratic discrimination (QDF), can be doubtful. Various procedures have been discussed as alternatives to classical discriminant analysis. Some of them are: regularized discrimination (RLDF or RQDF), kernel discriminant function (KDF), k nearest neighbors discrimination (kNN) and Naive Bayesian discrimination (NB) (Hand et al., 2001a, 2001b).

Currently, researchers in classification tend to combine procedures, based on similar types or different base classifiers (Kotsiantis et al., 2007; Rokach, 2009, 2010a, 2010b). Specifically, considerable attention has been paid lately to families of classifiers originating from two ideas: bootstrap aggregations and boosting. From the first group, we choose the classical bootstrap ensembles called bagging, while from the second, adaptive boosting (AdaBoost) is selected. Because these fusion procedures are time consuming, the constituent classifiers with high levels of complexity may cause computational problems. Combining may improve perfor-

mance of a simple, though not optimal, constituent classifier. To some extent, fusion has the possibility to circumvent the drawbacks of such a classifier.

Reduction of dimensionality may be achieved by both extraction (creating new variables, representing the discriminant properties of the original data set) and by selection, where a smaller subset of original variables is looked for. However, original variables are clear and familiar for a physician. The possibility of giving the physician an easily interpretable graph, which could support solving classification problems, was studied. A simple interpretation of two or three-dimensional graphs prompted verification as to whether discrimination based on a few original highest discriminating variables allows for the proper prediction of new subjects. Interpretation of such low-dimensional plots would not demand that physicians have specialized statistical knowledge or software.

The research presented herein is dedicated both to visualization of discrimination procedures and to their evaluation. The aim of the work was to perform a performance comparison and visualization, on low-dimensional plots, of parametric and nonparametric discriminant procedures for leukemia differentiation. This was achieved using gene expression data presented by Golub et al. (1999). The data set was obtained from human acute leukemia patients. Golub et al. (1999) looked for class discovery (the cluster analysis) and examined a special case of discriminant analysis.

Dataset and Reduction of Dimensionality

A leukemia data set was examined (Table 1). The material, used in the discriminant analysis, comes from Golub et al. (1999). The issue is the discrimination between two types of leukemia: acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). The data set contains gene expression data from Affymetrix U95 microarrays of 11 AML and 27 ALL patients.

Table 1. Characterization of applied medical data set

Data Set	Medical decision problem, coming from:		Number of cases	Variables number	Number of groups
Leukemia	Discrimination between AML and ALL	St. Jude Children's Research Hospital	38	3571	2

The first issue that often appears in discriminant analysis is the need to decrease the dimensionality of a problem without an essential loss of information from the data set. For this analysis, the statistical reduction of the number of variables was achieved with the usage of multivariate estimators. For discrimination, the minimizing Wilks' statistic (named for Samuel S. Wilks) and maximizing Mahalanobis distance between centroids of populations are equivalent. These statistics measure variability among groups, and so can be used for the selection of variables with the highest discriminative power.

Squared Mahalanobis distance in the p-dimensional space $(\mu_i - \mu_j)'$ $\Sigma^{-1}(\mu_i - \mu_j)$ (Duda et al., 2001, p. 107) is the measure of distance between two multivariate normal distributions $N(\mu_i, \Sigma)$ and $N(\mu_j, \Sigma)$ of equal covariance matrices Σ, where μ_i is the mean vector of the distribution in population π_i $(i, j = 1, \ldots, k)$. Analogously, squared Mahalanobis distance between observations x and y is defined as (by choosing the inverse of within-population covariance matrix as Q matrix, i.e. $Q = \Sigma_i^{-1}$ in the formula given by Webb, 2002 on p. 422):

$$r_i^2(x, y) = (x - y)'\Sigma_i^{-1}(x - y) \tag{1}$$

where Σ_i is the covariance matrix in the population π_i $(i = 1, \ldots, k)$. A point y can also represent the whole population π_i $(i = 1, \ldots, k)$, when $y = \mu_i$. Then, we obtain the squared Mahalanobis distance of point $\boldsymbol{x}$ from population π_i, defined as the squared Mahalanobis distance from $\boldsymbol{x}$ to the mean μ_i (Duda et al., 2001, p. 35 or p. 626):

$$d_i^2(x) = (x - \mu_i)'\Sigma_i^{-1}(x - \mu_i) \tag{2}$$

Mahalanobis distance is a fundamental distance useful in multivariate statistical problems. The advantage of Mahalanobis distance over Euclidean distance is that the former does not depend on the axes' units, because the p-dimensional vector of variables is "scaled" by the variances. Furthermore, it uses multivariate statistical dependencies (covariance in Σ_i). Euclidean distance is a special case of Mahalanobis distance (when all pairs of variables are uncorrelated and all variables have the unit standard deviation). Squared Mahalanobis distance may also be generalized into a so-called generalized squared distance (which is in fact dissimilarity, not metric) expressed as: $(x - \mu_i)'\Sigma_i^{-1}(x - \mu_i) + \ln(\det \Sigma_i)$. It incorporates an additional component equal to $\ln(\det \Sigma_i)$, connected with a measure of variability in the population π_i.

Selection of variables can be achieved with the application of a procedure that optimizes the corresponding selection criterion. To choose variables with the highest discriminative power, one can use different statistical criteria for entry or removal of variables, such as maximizing the smallest F ratio between pairs of groups or minimizing the overall Wilks' lambda. The number of variables in the final model, chosen by stepwise selection methods, depends on two parameters: the F-value for the change (including or removing the variable) in Wilks' lambda and the tolerance level for entering the variable (the value 10^{-3} was chosen). The F-value statistics measuring the significance of the change in Wilks' lambda, when a $(p+1)$-th variable is added to the model, is given as (Norusis et al., 1990, p. B–19, formula 1.26a):

$$F_{change} = \frac{n-k-p}{k-1} \cdot \left(\frac{\Lambda_p}{\Lambda_{p+1}} - 1 \right) \tag{3}$$

where F_{change} is interpreted as the F value for including variables or excluding them, k is the number of populations, p is the number of variables, n is the total number of cases in all k populations and Λ_p is the Wilks' statistic for the set of p selected variables.

For the nonparametric Bayesian classifier and other nonparametric classifiers, variables are selected on the basis of the criterion that minimizes the leave-one-out error of the quick first nearest neighbor classifier. Both methods allow one to obtain the same two or three most discriminating genes. Next, the subsets of 2, 3, 5 and 10 genes are applied in the construction of discriminant functions.

Description of Discriminant Methods

On the basis of the selected variables, various parametric and nonparametric discriminant methods were applied. The parametric class includes Bayesian linear and quadratic discriminant functions with different modifications, e.g. regularization (McLachlan, 2004). Parametric methods were examined for special cases, such as regularized discrimination (Duda et al., 2001; Webb, 2002), diagonal linear and diagonal quadratic discrimination (DLDF, DQDF), and Euclidean discrimination, i.e. nearest mean classifiers (NMC, identifies the pattern as belonging to the group with the closest centroid). All of these special cases are connected with a probabilistic distance, either Mahalanobis or generalized. Additionally, a selection of variables for parametric discriminant functions is connected with this

probabilistic distance, as mentioned previously. Some examined parametric classifiers were also applied in fusion with resampling methods.

Bayesian Parametric Discriminant Analysis. The main concern of discriminant analysis is connected with the second step, i.e. classification (identification). Let us assume the loss function with equal costs, when the observation is allocated to the population π_i, though in fact it comes from population π_j $(i, j = 1, \ldots, k)$. A posteriori probability is the probability that the respective case belongs to a specific population π_i $(i = 1, \ldots, k)$. This probability is based on our knowledge of the variables' values in disjoint populations $\pi_1, \ldots, \pi_k$ and on a priori probabilities. Bayesian discriminant methods compute a posteriori probability $p(\pi_i|x)$ of x belonging to population π_i. It is defined by applying Bayes theorem (Webb, 2002, p. 7) and after substituting $f(x)$ by $\sum_{j=1}^{k} q_j f_j(x)$ (from the law of total probability) as the following:

$$p(\pi_i|x) = \frac{q_i f_i(x)}{\sum_{j=1}^{k} q_j f_j(x)}; \quad i = 1, \ldots, k \tag{4}$$

where $q_i = p(\pi_i)$ is a priori probability of the population π_i and $f_i(x) = p(x|\pi_i)$ is the class-conditional probability density function in π_i. In the current work, a priori classification probabilities that a case belongs to the population were taken as proportional to sizes of groups ($q_1 = 11/38$ i $q_2 = 27/38$). Parametric methods assume that each population has known distribution, for example p-variate normal distribution $N(\mu_i, \Sigma_i)$, where the density f_i is given as (Rao, 1973, p. 575):

$$f_i(x) = 2\pi^{(-0.5p)}[\det(\Sigma_i^{-1})]^{(-0.5)} \exp\left[-0.5(x - \mu_i)'\Sigma_i^{-1}(x - \mu_i)\right] \tag{5}$$

On the other hand, the nonparametric approach is based on nonparametric estimates of group-specific probability densities f_i, for example using the kernel or nearest neighbor methods. The observation is classified to the group with maximum a posteriori classification probability. Thus, all Bayesian discriminant methods define the partition of the multidimensional space into disjoint classification regions, corresponding to populations $\pi_1, \ldots, \pi_k$.

Assuming multivariate normal distribution $N(\mu_i, \Sigma_i)$ in population π_i, we obtain discriminant rules based on squared Mahalanobis distance (with $\Sigma = \Sigma_i$ in formula (2)) or its generalization, given as the following:

$$D_i^2(x) = (x - \mu_i)'\Sigma_i^{-1}(x - \mu_i) + \ln\det(\Sigma_i) - 2\ln qi \tag{6}$$

where q_i is a priori probability of the population π_i, Σ_i is the covariance matrix and $\boldsymbol{\mu}_i$ is the mean vector for the population π_i. This formula defines a measure of "generalized squared distance".

The above-stated distances (2) and (6) are strictly connected with classification. Each element is classified as belonging to the group to which it is the closest in terms of distance (6). Using the generalized squared distance D_i^2, we can derive a posteriori classification probabilities as the following (SAS/STAT, 1990, p. 680):

$$p(\pi_i|x) = \frac{\exp(-0.5D_i(x))}{\sum_{j=1}^{k} \exp(-0.5D_j(x))}; \quad i = 1, \ldots, k \tag{7}$$

The probability (7) can be expressed using the quadratic discriminant score as well. Let us consider the classification function obtained from the distance $D_i^2(x)$, given as:

$$E_i(x) = -\tfrac{1}{2} D_i^2(x)$$

which can be rewritten as quadratic discriminant score (QDF) (Krzyśko, 1990, p. 15; Rao, 1973, p. 575):

$$E_i(x) = -\tfrac{1}{2}(x - \mu_i)' \Sigma_i^{-1} (x - \mu_i) - \tfrac{1}{2} \ln \det(\Sigma_i) + \ln q_i \tag{8}$$

$(i = 1, \ldots, k)$. For equal covariance matrices $\Sigma_i = \Sigma$ $(i = 1, \ldots, k)$, this classification function reduces to the formula of the linear classification function (LDF) (Krzyśko, 1990, p. 19; Rao, 1973, p. 575):

$$e_i(x) = (x - \tfrac{1}{2}\mu_i)' \Sigma_i^{-1} \mu_i + \ln q_i; \quad i = 1, \ldots, k \tag{9}$$

We can as well express the linear classification function in equivalent classical form (Krzyśko, 1990, p. 20):

$$e_i(x) = \mu_i' \Sigma_i^{-1} x - \tfrac{1}{2} \mu_i' \Sigma^{-1} \mu_i + \ln q_i; \quad i = 1, \ldots, k \tag{10}$$

and also as dependent on Mahalanobis distance:

$$e_i(x) = -0.5 d_i^2(x) + \ln(q_i); \quad i = 1, \ldots, k \tag{11}$$

In these formulas, the subscript i denotes the relevant group; $\boldsymbol{x}$ is the observed vector value. $E_i(x)$ and $e_i(x)$ are the resulting classification scores for observation x. Moreover, it holds:

$$p(\pi_i|x) = \frac{\exp(E_i(x))}{\sum_{j=1}^{k} \exp(E_j(x))}; \quad i = 1, \ldots, k \tag{12}$$

Thus, minimizing the distance $D_i^2(x)$ is equivalent to maximizing the discriminant score $E_i(x)$. In the case of $\Sigma_i = \Sigma_j = \Sigma$ $(i, j = 1, \ldots, k)$, it is also equivalent to maximizing $e_i(x)$. Moreover, taking the additional assumption of equal a priori probabilities $q_i = q_j$ $(i, j = 1, \ldots, k)$ into account, it is equivalent to minimizing the Mahalanobis distance $d_i(x)$. Linear or quadratic classification functions can be used to predict to which group each case most likely belongs. Each function permits us to compute classification scores for each case and for each group ((8)–(11)).

Both linear and quadratic discrimination methods assume that the data come from a multivariate normal distribution. However, deviations from the multivariate normality are usually not fateful. Additionally, for linear discrimination, the homoscedasticity assumption (that Σ_j are homogeneous across groups) is taken. Again, minor departures are usually not meaningful. Quadratic or linear discrimination with diagonal covariance matrices Σ_i, in which independence of variables in each discriminated group Π_i is assumed, is called diagonal or uncorrelated discrimination (DQDF, DLDF, for linear or quadratic, respectively). DQDF and DLDF have the advantage over classical classifiers in those situations in which covariance matrix singularity raises a problem. The DQDF method, in opposite to the DLDF method, includes different variances in groups, though like DLDF, it assumes linear independence of variables.

For the analysis conducted for this study, regularized QDF and LDF (when covariance matrices in populations are equal) were applied as single and then combined classifiers. Generally, regularization means avoiding overfitting to the training set. It uses penalization for the fit. Regularization for discrimination (RFD) was proposed by Friedman (2001). RDF is the modification of linear or quadratic classical classifiers (Pękalska, 2005, p. 94; Webb, 2002, p. 37). For $p >> N$, when the data matrix can be large, it has a rank of, at most, $N < p$. If then, the pooled covariance matrix Σ or particular matrices Σ_i, Σ_j become singular, the inverse cannot be derived. Then, a solution may be the usage of, e.g. the regularized form of covariance matrix Σ_i (Heijden et al., 2004): $G_i^{(r,s)} = (1 - r - s)\Sigma_i + r\,\mathrm{diag}(\Sigma_i) + s\,mI$, where r, s from the interval $\langle 0, 1\rangle$ are regularization parameters, $\mathrm{diag}(\Sigma_i)$ is the diagonal matrix obtained from Σ_i, m is the average value of p diagonal values in matrix Σ_i and I is the identity matrix. Regularized covariance matrix $G_i^{(r,s)}$ may be the base of a distance corresponding to Mahalanobis distance. For $r = s = 0$, we obtain a special case, which is a classical linear

or quadratic discriminant function. For $s = 0$, a version of the regularized discrimination that shrinks the covariance matrix towards its diagonal is obtained.

Nonparametric Methods. Bayesian kernel discriminant functions (KDF) and k-nearest neighbor classifiers (1NN or kNN) belong to the nonparametric class. Nonparametric Bayesian methods do not assume any form of distribution in discriminated populations. Thus, they use nonparametric estimates, such as the Parzen-Rosenblatt kernel method (Webb, 2002), for probability densities in all populations. For nonparametric Bayesian methods, either Mahalanobis distance or Euclidean distance can be employed. This distance from the given point x_0 is defined by matrix V_i (SAS/STAT, 1990, p. 681):

$$d_i^2(x, x_0) = (x - x_0)' V_i^{-1} (x - x_0) \tag{13}$$

where the matrix V_i ($i = 1, \ldots, k$) can be the pooled or within-groups matrix of covariances. To define the proximity (a posteriori probability), the kernel method uses a p-dimensional ellipsoid. The volume of this ellipsoid depends on the smoothing parameter (radius r for kernel method) and on the distance (13). Large r values produce more regular estimates of the density function; however, then large data sets are needed. The kernel method with radius r and matrix V_i can use different kernel functions, for example normal or uniform. The base of these kernel functions' definition is the volume of the ellipsoid defined by matrix V_i: $\{x : d_i^2(x) = r\}$, where the distance d_i is specified by formula (13). For the given kernel and the given radius r (equal for different groups), these volumes differ among groups if the covariance matrices are not equal in all k populations (SAS/STAT, 2008). In Bayesian kernel discrimination, the problem of correct estimation methods often occurs if the number of dimensions is high, relative to group sizes. In the current paper, the parameter r was chosen to minimize the criterion, which was the CV estimate of error rates.

The nearest-neighbor (NN) method uses the given number, k, of training set elements for each discriminated observation. The classifier finds the value (radius) that is based on the distance values from the classified observation to the k-th-nearest data point according to the chosen distance function (e.g. Euclidean or Mahalanobis distance). Classified observation is identified as belonging to the population associated with the training element that achieves the smallest (for 1-NN) or k-th smallest (for k-NN) distance function. Usually, selection of k is not crucial; however, choosing

values from 3 to 7 neighbors is preferred to prevent the possibility of ties. The k parameter was chosen in the presented investigation according to distance, as proposed by Lissack et al. (1976). The applied kernel function is the classical Gaussian one.

Other nonparametric density estimation than that of the kernel method is applied by the Naive Bayesian procedure (Duda et al., 2001; Hand et al., 2001b). The predicted class is the one with maximum a posteriori probability. However, the assumption that each of the class densities is a product of marginal densities is taken. Thus, a conditional independence in each population class is expected. Strong assumption causes the estimated density to be much simpler than the real density. Consequently, the Naive Bayesian classifier is especially useful in highly dimensional problems, when probability estimation is difficult for relatively small samples.

Classifier Fusion Procedures. Randomly generated subsets of training data joined with combined classifiers built on them were also considered as part of the presented study. Bagging (Bootstrap AGGregatING), introduced by Breiman (1996), is an ensemble based on bootstrap samples. It is created by randomly choosing samples n times from the learning set with possible replacements, where n denotes the size of the learning set. The classifier is trained on each bootstrap subsample. Resulting classifiers are then joined e.g. by averaging a posteriori probability or by taking the unweighted majority vote.

Boosting can also be considered as a method based on resampling of the learning data set. However, in boosting, selection of subsequent subsamples depends on the results of combined classifier performance achieved in previous loops. In sequentially generated learning sets, the weights of misclassified cases are increased, so that the ensemble creates the boosted, improved classifiers. In spite of the similarity in the step of resampling the training sets, bagging and boosting are different combining methods. However, both of them can help in the stability of the constituent classifier. The known examples of unstable learners are classification trees and neural networks.

"Boosting" (Freund et al., 1997) the performance of weak classifiers is connected with ARCing-Adaptive Resampling and Combining (Breiman, 1998). An important property of boosting is the resistance to outliers. Friedman (2001) proposed a few explanations for the resistance against overfitting in boosting procedures. The most popular boosting method is Adaptive Boosting (AdaBoost) (Freund et al., 1999). This procedure allows the designer to continue adding classifiers until some desired low

training error is achieved. AdaBoost is proficient in reducing training errors exponentially (Freund et al., 1997) and may realize boosting by resampling.

Freund et al. (1998) gave the bound for generalization errors and the theoretical analysis of any voting methods, so the bound is appropriate for bagging and boosting. The components of the bound consist of the training error and, additionally, of the confidence. This confidence is a decreasing function of the number of observations in the training set and is an increasing function of the "complexity" of the constituent classifiers, but the confidence does not depend explicitly on the number of constituent classifiers.

The bagging procedure is generally a variance reduction tool (Hastie et al., 2001). On the other hand, boosting methods mainly decrease the bias of the base procedure, though they may also reduce variance. Bagging and boosting ensembles can be applied theoretically to each classifier, though the limitations of time and memory exist. For bagging and boosting, decision trees are usually chosen as constituent classifiers (Breiman et al., 1984; Rokach et al., 2005). In the current paper, other classifiers, such as LDF, RLDF, QDF and NMC, were examined.

Classification Error Evaluation. Appropriate classification error estimation is a very important issue, although in cases where the amount of data is relatively small, the estimation of error rate is not straightforward. To compare the used methods' results, the proportion of misclassified cases was estimated. The estimate of error rates is unbiased when the test set is independent of the training set. However, to obtain such a holdout error-rate estimate, a large amount of data is needed. Thus, frequently, the expected loss for a new trial is estimated by cross-validation (CV) or leaving-one-out (L1o) techniques, which make better use of the data set. Applied error CV and L1o estimates (Ambroise et al., 2002) make use of the data set including only the elements applied to derive the discriminant criterion (training data set). The L1o method achieves a nearly unbiased error-rate estimate for the new, independent testing set. Performance is defined as $1 - e$, where e is error of classification.

A summary of the prediction, including a table of errors made for comparison purposes, tells how well the current discriminant functions predict group membership of cases. A comparison was made between different fusion methods of discriminant analysis and how they worked on the genomic data set. For this purpose, the resubstitution (apparent error), leaving-one out (L1o), and cross-validation (CV) estimates of the error-rates were used. The numerical results of single and combined classifiers are illustrated in the next part of the article, as a part of the visualization discussion.

Visualization of Single Parametric and Nonparametric Discriminant Functions

A leukemia data set is one for which we can find only two or three excellent genes that have satisfying discriminant power for classification (Xiong et al., 2001). For example, the two best discriminating variables named in the data set presented here are the genes M27891 and X04145, which are denoted on the presented graphs by the names "Gene2" and "Gene6". Thus, these genes can be useful for visualization of classical and nonparametric discriminant methods, for numerical and graphical comparison purposes.

In all Bayesian discriminations related to the presented figures, unequal a priori probabilities proportional to group sizes were taken. Assuming p-variate normal distribution $N(\mu_i, \Sigma_i)$ for i-th population $(i = 1, \ldots, k)$, the decision boundary between each pair of populations π_i and π_j is defined by a quadratic equation

$$\{x : E_i(x) = E_j(x)\}$$

for $(i, j = 1, \ldots, k)$. This equation defines $(p-1)$ dimensional quadric boundary surface between populations π_i and π_j (Figure 1). In practice, we do not know the parameters of the normal distributions, so they are estimated using the training data.

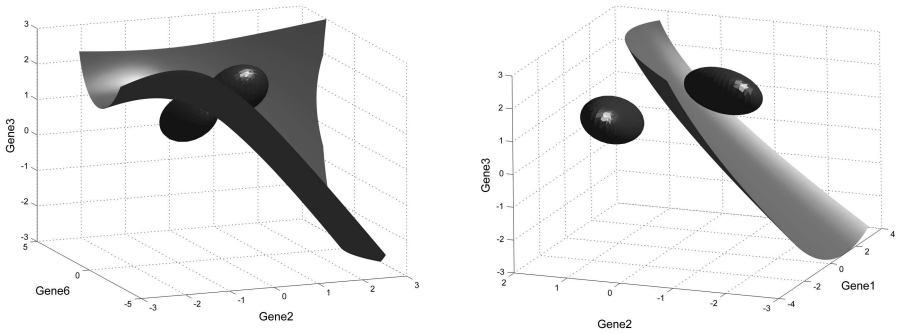

Figure 1. Ellipsoids of estimated a posteriori probabilities with quadric classification boundary equalizing generalized distances (6) from population centroids, for variables x_2, x_6, x_3 (left) and x_1, x_2, x_3 (right); apparent errors = 0.105, CV = 0.12 (SD = 0.13)

The quadric discriminating each pair of populations π_i (probability density is represented by ellipsoids, Figure 1) is obtained for the classical Bayesian classifier. For unequal covariance matrices Σ_i and Σ_j $(i, j =$

$1, \ldots, k$), the kind of quadric discriminating each pair of populations depends on whether the matrix $\Sigma_i^{-1} - \Sigma_j^{-1}$ (or $\Sigma_j^{-1} - \Sigma_i^{-1}$) is a positive definite matrix (then a hyper-ellipsoidal boundary is obtained) or is a semi-positive definite one (ellipsoidal cylinder, parabolic-type cylinder or ellipsoidal paraboloid are possible). Then, the eigenvectors corresponding to nonzero eigenvalues of the above stated positive definite (or semi-positive definite) matrix determine the hyper-ellipsoid axes (Krzyśko, 1974). If matrix $\Sigma_i^{-1} - \Sigma_j^{-1}$ (or $\Sigma_j^{-1} - \Sigma_i^{-1}$) is not defined as positive or semi-positive, one obtains, for example, a hyperboloid. For multivariate normal distribution $N(\mu_i, \Sigma_i)$, the concentration hypersurfaces of the density create a family of hyper-ellipsoids with the common center point μ_i (Morrison, 1990). Hyper-ellipsoid axes' lengths are given by the eigenvectors of the covariance matrix. The lengths of the succeeding axes are proportional to the square root of non-increasing eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ of the covariance matrix Σ_i. For quadratic discrimination, the subsets of p-dimensional space that are classification regions are not necessarily the joined area, if the data set is more complex. Then, even for the same classification region, we may obtain separated parts. In two-dimensional space, this case corresponds to two branches of a hyperbolic boundary. In three-dimensional space, such boundaries can be, for example, hyperboloids. Two and three-dimensional examples of Bayesian discrimination in Figures 1–6 present several different forms.

Three-dimensional Gaussian distribution yields Bayesian decision boundaries that are either hyperplanes or hyperquadrics. In the case of three-dimensional Gaussian discrimination, the quadric surface separates quadratic decision regions (Figure 1). Taking $p = 3$ selected the most discriminated variables (Genes 2, 3 and 6, i.e. M27891, X04145 and U05259); the ellipsoids around centroids of populations μ_1 and μ_2 (Figure 1) are obtained as surfaces for an arbitrary a posteriori probability $p(\pi_i|(x)$ (7) or generalized distance D_i (6). The three-dimensional surface of constant probability densities in each discriminated population π_i also creates an ellipsoid. The shape of the ellipsoid is the same because it is also determined by the inversed covariance matrix Σ_i^{-1}.

The two ellipsoids for a chosen value of a posteriori probability intersect in a boundary, which generally is a quadric. If the densities are more complicated than Gaussian, the classification boundary regions may be more complex.

On the left side of Figure 1, the boundary is constructed on Gene2, Gene6 and Gene3. Matrix $\Sigma_i^{-1} - \Sigma_j^{-1}$ is not defined as positive (negative) or semi-positive (semi-negative),

$$\begin{matrix} 3.5758 & 0.4104 & 0.7807 \\ 0.4104 & -3.0298 & 1.3152 \\ 0.7807 & 1.3152 & 2.7481 \end{matrix}$$

so the boundary is hyperboloid. The corresponding equations for classifiers (8) for two groups, respectively, are calculated as:

$$E_1(x) = (\, x_2 \quad x_6 \quad x_3 \,) \begin{pmatrix} -2.4328 & -0.4778 & -0.8815 \\ -0.4778 & -6.3347 & 0.6064 \\ -0.8815 & 0.6064 & -4.8122 \end{pmatrix} (\, x_2 \quad x_6 \quad x_3 \,)' +$$

$$+ (\, -1.2930 \quad 1.6206 \quad 16.4592 \,)(\, x_2 \quad x_6 \quad x_3 \,)' - 4.7456$$

$$E_2(x) = (\, x_2 \quad x_6 \quad x_3 \,) \begin{pmatrix} -0.6450 & -0.2728 & -0.4911 \\ -0.2728 & -7.8508 & 1.2643 \\ -0.4911 & 1.2643 & -3.4381 \end{pmatrix} (\, x_2 \quad x_6 \quad x_3 \,)' +$$

$$+ (\, 1.8428 \quad -9.7621 \quad 8.3512 \,)(\, x_2 \quad x_6 \quad x_3 \,)' - 4.7456$$

The identification of x is carried out for the first group if $E_1(x) > E_2(x)$. If not, it is carried out for the second group. For comparison of quadrics in three-dimensional space, Gene1, Gene2 and Gene3 are also applied. On the right side of Figure 1, the quadric boundary for Gene1, Gene2 and Gene3 corresponds to two 3×3-covariance matrices Σ_1 and Σ_2, i.e.

$$\begin{matrix} 0.1458 & 0.0814 & 0.0000 \\ 0.0814 & 0.2256 & -0.0440 \\ 0.0000 & -0.0440 & 0.1138 \end{matrix}$$

and

$$\begin{matrix} 0.0902 & 0.0595 & 0.0121 \\ 0.0595 & 0.9141 & -0.1512 \\ 0.0121 & -0.1512 & 0.1796 \end{matrix}$$

Matrix $\Sigma_1^{-1} - \Sigma_2^{-1}$ is now defined as positive, so the classification boundary is designed from the ellipsoid (Figure 1, right). Two separate classification regions are single joint regions. In Figure 1, only parts of the classification boundaries are visible. The misclassification rates, corresponding to the right and left parts of Figure 1, are the same, according to CV and L1o (presented in Table 2 for best three genes: Gene6, Gene2 and Gene3). Performance results for discriminant functions, illustrated in Figure 1, as well as for other single classifiers built on the best three genes, are presented in Table 2, in columns for 3 variables (CV and L1o).

Table 2. Generalization errors (*100%) of single classifiers for 2, 3, 5 and 10 selected genes

nr of variables	2		3		5		10	
Method/error	CV	L1o	CV	L1o	CV	L1o	CV	L1o
LDF	10.5	10.5	7.9	7.9	7.9	7.9	8.4	7.9
QDF	10.0	10.5	11.6	10.5	5.5	5.3	19.7	18.4
RLDF $r = 0.2$; $s = 0.5$	10.5	10.5	7.9	7.9	8.2	7.9	5.3	5.3
RQDF $r = 0.2$; $s = 0.5$	10.5	10.5	10.5	10.5	7.9	7.9	5.3	5.3
RLDF $r = 0.02$; $s = 0.05$	10.5	10.5	8.2	7.9	7.9	7.9	7.6	5.3
RQDF $r = 0.02$; $s = 0.05$	10.5	10.5	10.5	10.5	6.8	7.9	5.8	7.9
DQDF	10.5	10.5	10.5	10.5	7.9	7.9	5.8	5.3
Parzen	10.5	10.5	10.3	7.9	7.9	5.3	5.3	5.3
Parzen $r = 1$	9.0	10.5	7.4	7.9	7.9	7.9	5.5	5.3
kNN	8.4	7.9	7.9	7.9	7.6	7.9	5.8	5.3
1NN	6.3	5.3	2.6	2.6	2.9	2.6	7.9	7.9
3NN	14.2	15.8	12.9	13.2	7.9	7.9	5.3	5.3
5NN	10.5	10.5	9.7	10.5	7.9	7.9	5.3	5.3
NB	10.5	10.5	8.2	7.9	8.2	7.9	5.0	5.3
min	6.3	5.3	2.6	2.6	2.9	2.6	5.0	5.3
max	14.2	15.8	12.9	13.2	8.2	7.9	19.7	18.4
mean	10.3	10.3	9.0	8.8	7.3	7.1	7.0	6.8
sd	1.7	2.2	2.5	2.4	1.4	1.6	3.8	3.5
var.coeff(%)	16	21	27	27	20	23	55	52

Underlined – best results

For linear discrimination LDF, we obtained the following three dimensional classifier functions:

$$e_1(x) = (\,0.1147 \quad 2.0317 \quad 14.8041\,)(\,x_2 \quad x_6 \quad x_3\,)' + 13.9228$$

$$e_2(x) = (\,0.1147 \quad 2.0317 \quad 14.8041\,)(\,x_2 \quad x_6 \quad x_3\,)' - 7.9932$$

The identification of a new pattern $\boldsymbol{x}$ is performed for the first group if $e_1(x) > e_2(x)$. If not, it is performed for the second group.

For the reason that in three-dimensional space it is difficult to analyze the details of the discriminant procedure from the inside, the next presented figures are illustrations of two-dimensional discrimination. *Eleven* AML patients are denoted on the plots by stars ("*") and 27 ALL patients by

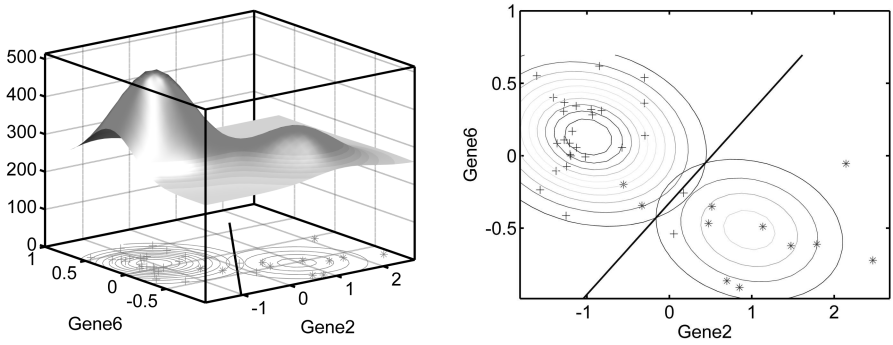

Figure 2. Estimated classifier functions and boundary of linear discrimination LDF; apparent error = 0.105, CV = 0.105 (SD = 0.07)

symbols "+". Apparent (resubstitution) errors and CV errors with corresponding standard deviations (SD) are given in the titles of the plots.

For a linear classification, equal covariance matrices are assumed. In consequence, linear boundaries are obtained (Figure 2). If a discrimination is performed on more than two classes, the decision boundary for any pair of populations is a hyperplane in p-dimensional space, so all pair decision boundaries are linear. The LDF misclassification error is connected with the area of overlap between the two densities in discriminated populations. For different a priori probabilities $q_i \neq q_j$, the Bayesian classification error of LDF is equal to (Krzyśko et al., 2008, p. 33):

$$q_1/(q_1+q_2)F(-0.5M + M^{-1}\ln(q_2/q_1)) +$$

$$+\ q_2/(q_1+q_2)F(-0.5M - M^{-1}\ln(q_2/q_1))$$

where $M = \sqrt{(\mu_i-\mu_j)'\Sigma^{-1}(\mu_i-\mu_j)}$ is the Mahalanobis distance between two groups' centroids μ_i, μ_j and F is the cumulative distribution function of multidimensional normal distribution $N(\mu_i, \Sigma)$. Thus, assuming equal a priori probabilities, the Bayesian error of LDF is equal to $F(-0.5M)$ (Morrison, 1990, p. 348).

The classification boundary is a set of points equalizing Mahalanobis distance modified by $\ln(\det(\Sigma_i)) - 2\ln(q_i)$, given by formula (6). For equal covariance matrices, $\ln(\det(\Sigma_i))$ can be neglected. If equal a priori probabilities q_i and common Σ are assumed, then the classification boundary is fixed in the 'middle' between the projected means vectors, and each point of the boundary lies in equal Mahalanobis distance from the two populations' centroids μ_i and μ_j. If however the q_i were different in populations,

moving the cut-point toward the smaller population would improve the classifier performance (6). The Bayesian rule (7), incorporating different a priori probabilities q_i, minimizes the expected loss.

If q_i are not equal, they are usually estimated by proportions of observations belonging to the differentiated population. In this way, the Bayesian linear classification rule results in the shift of the boundary by $2\ln q_i$ toward the class with the smaller a priori probability. In the presented work, probabilities proportional to population sizes in the training set were taken, i.e. $q_1 = 11/38 = 0.289737$ and $q_2 = 27/38 = 0.710526$. In the examined dataset, q_1 and q_2 are apparently different, thus shifting improves the performance. Therefore, ellipses from LDF visualizing classification functions e_i have different values in corresponding contours. At the same time, the identical shape of contours and the same direction of axes is visible. This comes from assumption of equal covariance matrices in populations (or the same pooled Σ matrix), so both eigenvalues ($\lambda_1 \geq \lambda_2$) are also the same in both matrices.

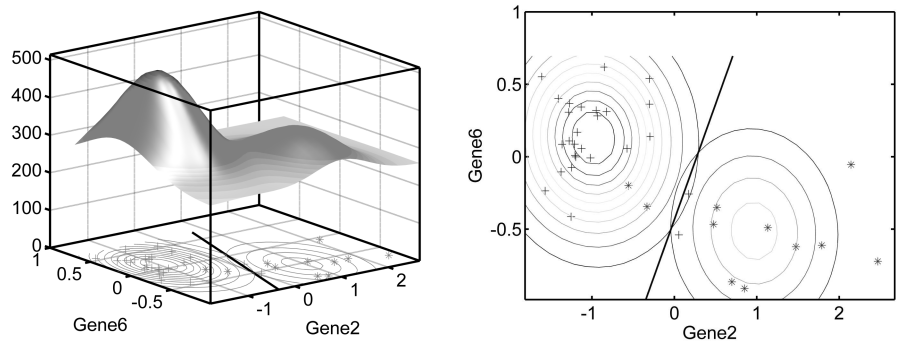

Figure 3. RLDF regularized with parameters $r = 0.2$ and $s = 0.5$; apparent error = 0.105, CV = 0.105 (SD = 0.01)

Though regularization with parameters $r = 0.2$ and $s = 0.5$ slightly changes the classification boundary of RLDF in comparison to LDF, the errors are not changed (Figure 3, Table 2). Those parameters do not apparently modify the linear classifier, but diminishing both parameters r and s could create regularized classification more different from LDF.

For quadratic QDF in Figure 4, the classification boundary is a set of points equalizing Mahalanobis distance modified by $\ln(\det(\Sigma_i)) - 2\ln(q_i)$, given by the formula (6). Component $\ln(\det(\Sigma_i))$ is a measure of diversity in population π_i ($i = 1, 2$). Thus, in the theoretical case of the equal remaining part: $(x - \mu_i)'\Sigma_i^{-1}(x - \mu_i) - 2\ln(q_i)$ (which was the criterion for linear classification of function LDF), another population might be identified. For

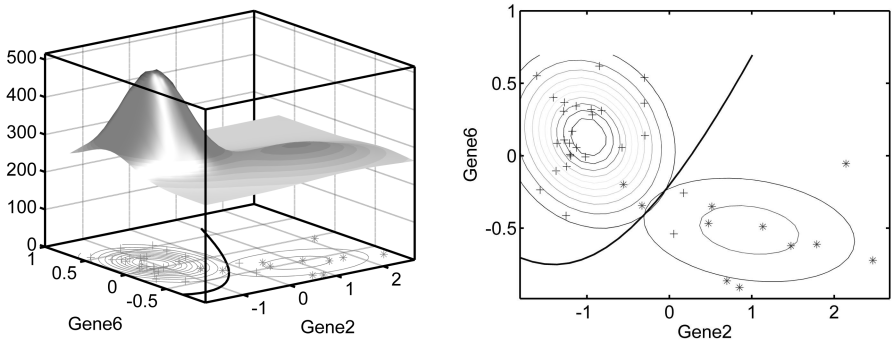

Figure 4. Estimated classifier for quadratic discrimination QDF; apparent error = 0.105, CV = 0.1 (SD = 0.01)

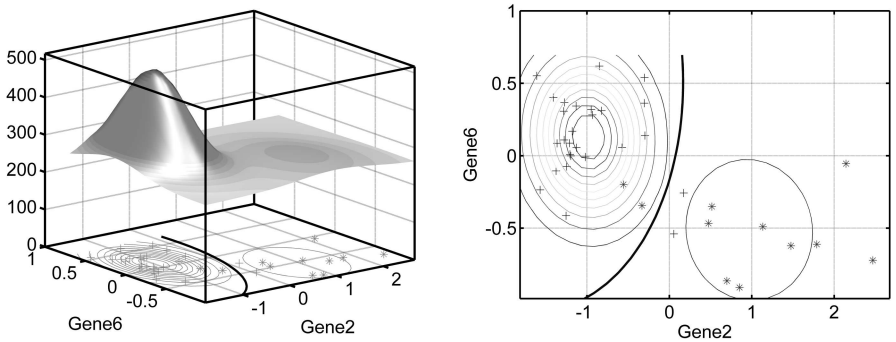

Figure 5. RQDF regularized with parameters $r = 0.2$ and $s = 0.5$; apparent error = 0.105, CV = 0.105 (SD < 0.001)

the resulting matrix $\Sigma_1^{-1} - \Sigma_2^{-1}$, which is not (semi) positive-definite, we obtained the hyperbola, visible as the boundary for quadratic discrimination in Figure 4.

The RQDF regularized classifier with parameters $r = 0.2$ and $s = 0.5$ (Figure 5) does not show apparent difference in the classifier boundary in comparison to the plot of classical QDF, and errors remain the same. Diminishing both parameters to $r = 0.02$ and $s = 0.05$ causes greater difference in the regularization matrix $G_i^{(r,s)}$, in comparison to covariance matrix Σ_i. However, the performance of the classifier, when based on two variables, was not improved for RQDF in comparison to QDF discrimination, but was improved for a larger number of genes (Table 2).

In Figure 6, a linear independence of two genes is assumed in each population. Thus, axes of classification function contours are orthogonal to coordinate axes. This is explained by the fact that eigenvectors of covariance

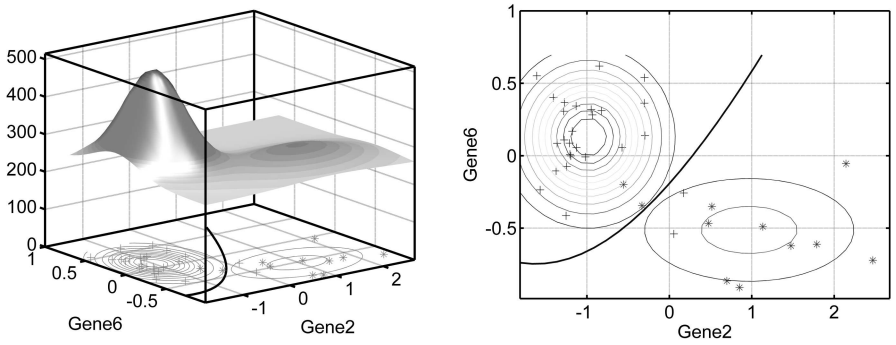

Figure 6. Uncorrelated (diagonal) DQDF; apparent error = 0.105, CV = 0.105 (SD < 0.001)

matrices are orthogonal to coordinate axes. Uncorrelated (diagonal) DLDF and DQDF are very useful for classification of high-dimensional issues in the bioinformatics domain, where we usually face singularity and problems with inversion of the matrices. However, in small dimensionality ($p = 2, 3, 5$ and 10 in our case), the apparent benefit over classical discrimination QDF is not obtained for less than 10 genes; for 2 genes, the misclassification rates are the same (Table 2).

A quite different boundary was obtained for the Naive Bayes model, where the assumption of normality is not made (Figure 7). For each of the discriminated populations, the Naive Bayesian model assumes the independent features x_l $(l = 1, \ldots, p)$, i.e. (Hand et al., 2001b, p. 353): $P((x_1, \ldots, x_p)|\pi_k) = P(x_1|\pi_k)P(x_2|\pi_k) \ldots P(x_p|\pi_k)$.

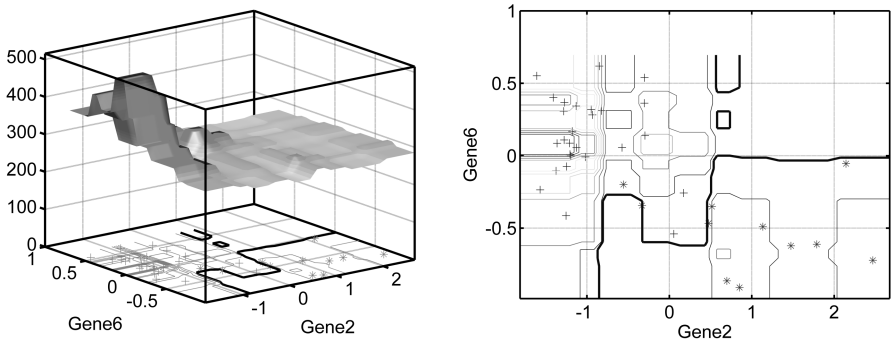

Figure 7. Density estimation and boundary of Naive Bayesian classifier; apparent error = 0.05, CV = 0.105 (SD < 0.001)

The nonparametric kernel classifier with Gaussian kernel and radius $r = 1$ is presented in Figure 8. Contours of density obtained by Parzen-

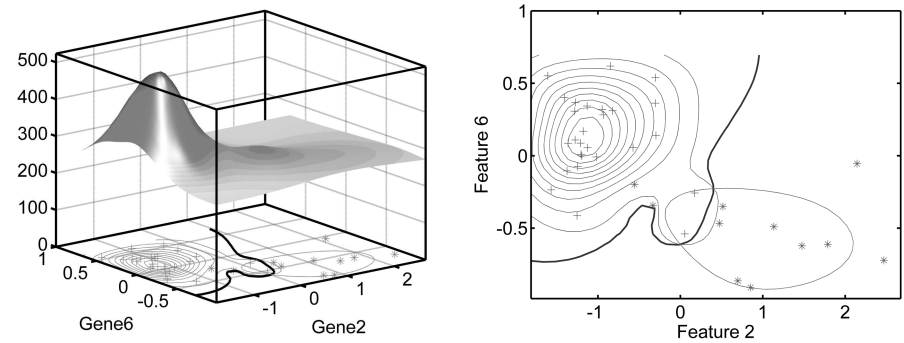

Figure 8. Gaussian kernel discriminant function with $r = 1$ – estimated densities and boundary error = 0.05, CV = 0.09 (SD = 0.04)

Rosenblatt kernel density estimation show that distributions in differentiated groups do not differ substantially from the two-dimensional Gaussian distributions; thus, classification error is similar to that of parametric classifiers. The boundary is then the most similar to that of quadratic discrimination, though it is more flexible for the part of the space with mixed classes in two-dimensional space. For Gene2 between -0.5 and 0.5 and Gene6 between -0.6 and -0.3, this boundary was closer to the Naive Bayes boundary. The shape of the nonparametric nearest neighbor classification boundary strongly depends on the number of neighbors. For the 1NN classification region for the AML group ("stars" marks), it consisted of two parts (Figure 9).

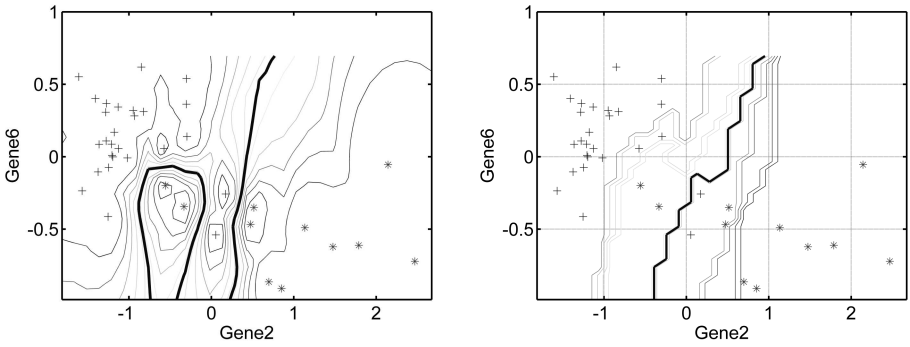

Figure 9. Nearest neighbor classifiers 1NN (apparent error = 0.105, CV = 0.06, SD = 0.03) and 5 kNN (apparent error = 0.105, CV = 0.105, (SD < 0.01)

Graphs made for visualization purposes were obtained by the author's own programs, based on the PRTOOLS package for Matlab (Duin et al., 2007).

Visualization of Combined Classifiers

Methods based on resampling of the training set, typically bagging and AdaBoost, were based on various constituent classifiers.

The NMC method, which can be regarded as a special case of classical LDF (for $\Sigma = I$), was applied to resampling ensembles. In addition, resampling fusion techniques were applied to the regularized linear and quadratic classifiers (RLDF and RQDF).

The bagging ensemble of classical LDF with 100 loops was compared to bagged regularized classifiers RLDF (Figure 10, left). The errors obtained by the resampling ensemble for both constituent classifiers were the same, i.e. CV = 0.105. The right side of Figure 10 contains classifiers obtained by 100 loops of bagging when constituent classifiers are QDF and regularized RQDF (in both cases CV = 0.105). The boundary function of bagged RQDF (tiny, dotted line) has a more regular shape than QDF (Figure 10, right). For discrimination between ALL and AML with the usage of the two best discriminating variables, bagging 100 loops was not beneficial to either QDF or RQDF.

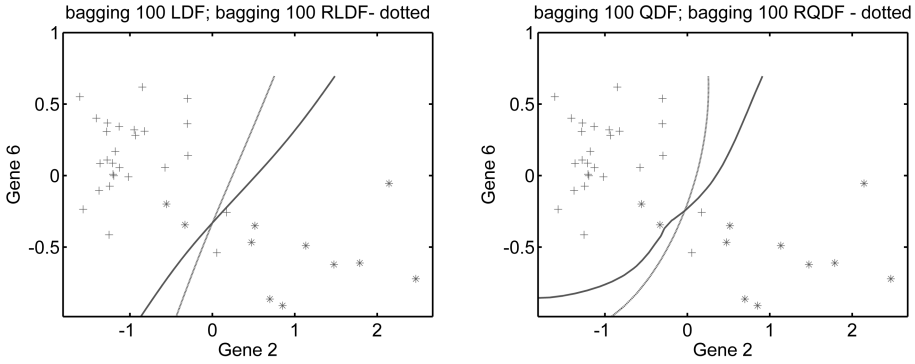

Figure 10. Final classification boundaries after 100 loops of bagging for linear LDF and RLDF (left and dotted: apparent, CV and L1o errors = 0.105) and QDF with RQDF (right and dotted: apparent error, CV = 0.105 with SD = 0.07, L1o = 0.18)

The adaptive boosting procedure AdaBoost was applied with a small number of loops (10) to allow presentation of combined classification boundaries for both constituent and merged classifiers. During the implementation of ten steps of boosting, after consecutive loops, the classifiers are focused on previously incorrectly classified objects. Ten boundaries, visible on the left and right sides of Figure 11, were obtained by use of this method. The merged classification boundary (Figure 11, bold lines) joins ten constituent

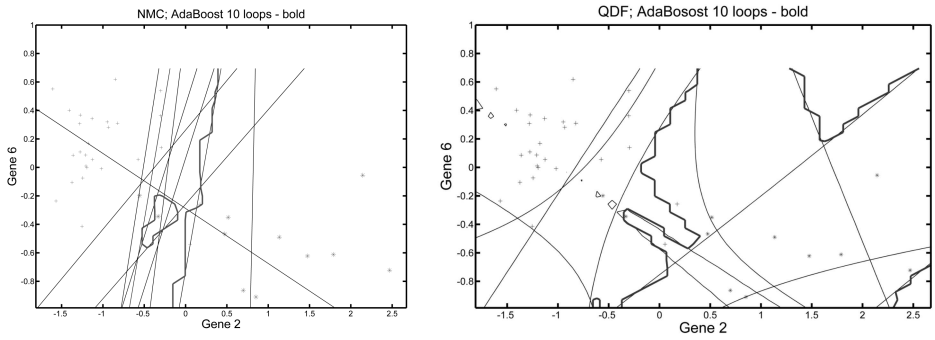

Figure 11. Ten loops with constituent and boosting boundaries. Left and bold: boosted Nearest Mean Classifier – Apparent err. = 0.05, CV = 0.105 (SD = 0.02), L1o = 0.105. Right and bold: boosted QDF-Apparent err. = 0.08, CV = 0.13 (SD = 0.05), L1o = 0.16

boundaries according to a voting scheme. According to cross-validation errors of single (Table 2) and combined NMC classifiers, ten bagging loops joined with the constituent linear NMC classifier were not beneficial to basic NMC classifiers (Figure 11, left). Similarly, for classical QDF, ten loops of boosting (Figure 11, right) were not preferred over the base generalization error of QDF (Table 2).

Similar errors in the case of classification do not necessarily mean that the same cases have the same classifications. In particular, erroneous classification to one of the groups may be more risky from a medical point of view. Thus, the analysis of conditional classification errors for each group, associated with specificity and sensitivity, might be an additional suggestion for assisting one in choosing which of the methods for solving a particular medical problem is more preferred. From this practical point of view, doctors may be guided by the criteria of sensitivity and specificity, which is close to their knowledge. Sensitivity and specificity may also be assessed by eye from plots. Apart from the assessment of apparent generalization errors, the classification boundary shape with overlaid scatter-plots of groups may also indicate whether the particular kind of classifier is overtrained or underfitted for the examined problem.

Discussion

Errors estimated by CV and L1o allow one to draw similar conclusions. The increase in numbers of variables from 2 to 10 leads to an improved performance for all examined classifiers, except for 1NN and QDF.

The QDF exception may be caused by the fact that the covariance matrix of 10×10 size was estimated on the basis of only 11 observations from a smaller group (AML). L1o error was not smaller for 2 than for 3 variables. In the same way, this also holds true for CV, again with the exception of QDF. The smallest CV errors for 2, 3 and 5 variables were achieved for 1NN discrimination. Many discriminant methods for 2 chosen variables achieved the same performance (CV and L1o error 10.5%). This is also visible by eye during comparison of Figures 2–9. Thus, the smallest variability coefficient of CV and L1o errors was obtained for 2 variables (16% and 21%, respectively), while the highest one was met for 10 variables (55% and 52%, respectively).

It is of interest to compare the performance obtained in the presented work to other authors' outcomes. Golub et al. (1999) applied a weighted gene voting method, which is a variant of a special case of linear discriminant analysis (DLDF). The authors of this work applied 50 selected genes and obtained a test error of 5.8%, which in comparison with errors of several applied methods (with 2, 3, 5 and 10 genes in Table 2), is smaller. However, some methods presented in Table 2 that are based on a small number of genes obtained better performance, for example 1NN with 2, 3 and 5 variables (L1o equal 5.3, 2.6 and 2.6, respectively).

Dudoit et al. (2002), using an apparently larger number of genes (50) for ALL and AML discrimination, obtained a test error smaller than Golub et al. (1999) by applying DLDF. However, LDF based on 50 genes obtained a higher test error (about 20%) than all classifiers presented in Table 2. Additionally, CART (tree) with a test error of about 10% apparently had the worst performance of all the methods, based on the small number of 5 and 10 genes from Table 2.

For a larger number of genes (200), Dettling (2004) obtained an estimate of error equal to 0.04 for single kNN. He also estimated error rates of 0.04 and 0.06, respectively, for advanced BagBoosting and boosting trees. For support vector machine (SVM), Ambroise et al. (2002) achieved (test, bootstrap and CV) errors of 0.15, 0.10, 0.07 and 0.05 for 4, 8, 32, and 210 genes, respectively. SVM for 4 genes was inferior to several methods based on 2 genes presented in Table 2, especially 1NN. This classifier performed similarly to many classifiers based on 3 and 5 genes, however for 210 genes it was better than most of the discriminations including 10 genes in Table 2.

Xiong et al.'s (2001) results indicated that the expression information from three or four genes was optimal for tumor classification in three data sets: leukemia, colon cancer and breast cancer. Additionally, as few as two genes achieved misclassification rates below 0.10. This is confirmed in the

presented paper. Thus, the two-dimensional plots for discriminant methods presented herein have obtained a motivation for practical usage.

Increasing the number of predictor variables from $p = 2$ to $p = 10$ improved the accuracy of the classifiers. The learners were comparable with results for apparently larger numbers of genes obtained by other authors.

Conclusions

For a small number of genes, the 1NN classifier achieved the smallest CV and L1o errors. The misclassification rates obtained for two variables and 1NN were comparable to various classifiers presented in other authors' works, for which apparently higher numbers of genes were applied.

Bivariate plots are illustrations of different types classifiers in the general multidimensional situation. In p-dimensional spaces, the lines are replaced by hyperplanes, and curves of second order on the plane are substituted by quadrics. However, for more complex nonlinear discriminant functions, the complex surfaces of boundaries are obtained.

The results present the possibility of cancer classification based on the monitoring of the gene expression of a small number of best discriminating variables. The outcomes can suggest an approach for recognizing other types of cancer, where only a few genes have a sufficient amount of discriminative power. By monitoring only a small subset of genes, the costs in medical diagnostics may be diminished.

Because the two chosen variables were well discriminating for the examined diagnostic problem, there were no significant differences between the errors for the majority of the analyzed methods, both for CV or L1o. However, for problems presented by more complex data sets, this difference may be more pronounced.

Resampling methods of combining, known as especially beneficial for unstable classifiers, did not appear beneficial to examined single classifiers. Linear classifiers may be unstable, which is observed for relatively small numbers of training cases in comparison to dimensionality (Skurichina, 2001). However, the examined data set connected with classification boundaries did not show properties of instability, especially for linear and quadratic classifiers. For the examined, well differentiable data set without noise, single classifiers created on the basis of two selected variables had good discriminant properties. In this way, the ensembles – bagging with a hundred loops and boosting with ten loops – obtained by resampling, were not beneficial.

REFERENCES

Ambroise, C., & McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences of the United States of America*, 99(10), 6562–6566.

Breiman, L. (1996). Bagging predictions. *Machine Learning*, 24(2), 123–140.

Breiman, L. (1998). Arcing classifiers. *The Annals of Statistics*, 26(3), 801–849.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Belmont: Wadsworth.

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

Duda, R. O., Hart P. E., & Stork, D. G. (2001). *Pattern Classification*. New York: Wiley & Sons.

Dudoit, S., Fridlyand, J., & Speed, T. P. (2002). Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association*, 97(457), 77–87.

Duin, R. P. W., Juszczak P., Paclik, P., Pekalska, E., de Ridder, D., Tax, D. M. J., & Verzakov, S. (2007). *PRTools 4.1. A Matlab Toolbox for Pattern Recognition*. Delft University of Technology.

Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5), 1189–1232.

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1), 119–139.

Freund, Y., & Schapire, R. E. (1998). Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods. *The Annals of Statistics*, 26(5), 1651–1686.

Freund, Y., & Schapire, R. E. (1999). A Short Introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5), 771–780,

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., et al. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286, 531–537.

Hand, D. J., & Yu, K. (2001a). Idiot's Bayes – not so stupid after all? *International Statistical Review*, 69(3), 385–398.

Hand, D. J., Mannila, H., & Smyth, P. (2001b). *Principles of data mining*. Massachusetts Institute of Technology.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning*. Springer, New York.

Heijden, F., Duin, R. P. W, de Ridder, D., & Tax, D. M. J. (2004). *Classification, Parameter Estimation and State Estimation. An Engineering Approach Using MATLAB*. England: Wiley.

Kotsiantis, S. B, Zaharakis, I. D., & Pintelas, P. E. (2007). Supervised machine learning: A review of classification techniques. *Artificial Intelligence Review*, 26(3), 159–190.

Krzyśko, M. (1974). Kwadratowe funkcje dyskryminacyjne. *Matematyka Stosowana*, II, 151–156.

Krzyśko, M. (1990). *Analiza dyskryminacyjna*. Warszawa: WNT.

Krzyśko, M., Wołyński, W., Górecki, T., & Skorzybut, M. (2008). *Systemy uczące się: rozpoznawanie wzorców, analiza skupień i redukcja wymiarowości*. Warszawa: WNT.

Lissack, T., & Fu, K. S. (1976). Error estimation in pattern recognition via L-distance between posterior density functions. *IEEE Transactions on Information Theory*, 22(1), 34–45.

Marchiori, E., & Sebag, M. (2005). Bayesian Learning with Local Support Vector Machines for Cancer Classification with Gene Expression Data. In F. Rotlauf et al. (Eds) Lecture Notes in Computer Science: Vol. 3449. Applications of Evolutionary Computing (pp. 74–83). Lausanne, Switzerland. Springer Verlag.

McLachlan, G. (2004). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley.

Morrison, D. F. (1990). *Multivariate Statistical Methods* (3rd ed.) (R. Zieliński, Trans.). New York: McGraw–Hill Book Company.

Norusis, M. J, & SPSS, Inc. (1990). *SPSS PC+ Advanced Statistics. Release 4.0*. Chicago: SPSS Inc.

Pękalska, E. (2005). *The Dissimilarity representations in pattern recognition. Concepts, theory and applications* (Doctoral thesis). ASCI Dissertation Series no. 109. Delft University of Technology. Delft, The Netherlands.

Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., & Sturla, L. M., Angelo, M., McLaughlin, M. E., Kim, J. Y., et al. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870), 436–442.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications* (2nd Ed.). New York: Wiley.

Rokach, L. (2009). Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography. *Computational Statistics and Data Analysis*, 53(12), 4046–4072.

Rokach, L. (2010a). Pattern Classification Using Ensemble Methods. In H. Bunke, & P. S. P. Wang (Eds.), *Series in Machine Perception and Artificial Intelligence* (Vol. 75). World Scientific Publishing.

Rokach, L. (2010b). Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1–2), 1–39.

Rokach, L., & Maimon, O. (2005). Top-down induction of decision trees classifiers – a survey. *IEEE Transactions on Systems, Man and Cybernetics*, *Part C: Applications and Reviews*, 35(4), 476–487.

SAS/STAT (1990). *User's Guide. Version 6.* Cary, NC, USA: SAS Institute Inc.

SAS/STAT (2008). *9.2 User's Guide.* Cary, NC, USA: SAS Institute Inc.

Skurichina, M. (2001). *Stabilizing weak classifiers* (PhD thesis). Delft University of Technology, Delft, The Netherlands.

Xiong, M., Li, W., Zhao, J., Jin, W., & Boerwinkle, E. (2001). Feature (Gene) Selection in Gene Expression-Based Tumor Classification. *Molecular Genetics and Metabolism*, 73(3), 239–247.

Webb, A. R. (2002). *Statistical pattern recognition.* England: Wiley.