Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners

Sarunas J. Raudys and Anil K. Jain, *Fellow, IEEE*

Abstract—**During the last two decades a considerable amount of effort has been devoted to the analysis of the influence of both training and testing sample size on the design and performance of pattern recognition systems. These questions are interesting to practitioners as well as theoreticians, because the small-sample effects can easily contaminate the design and evaluation of a proposed system. For applications with a large number of features and a complex classification rule, the training sample size must be quite large. A large test sample is required to accurately evaluate a classifier with a low error rate. The design of a pattern recognition system consists of several stages: data collection, formation of the pattern classes, feature selection, specification of the classification algorithm, and estimation of the classification error. In this paper, we will discuss the effects of sample size on feature selection and error estimation for several types of classifier. In addition to surveying prior work in this area, our emphasis is on giving practical advice to today's designers and users of statistical pattern recognition systems.**

Index Terms— **Classification error, classifier design, curse of dimensionality, feature selection, statistical pattern recognition, test samples, training samples.**

I. Introduction

DURING the last two decades a considerable amount of effort has been devoted to the analysis of the influence of both training (also called design or learning) and testing sample size on the design and performance of pattern recognition systems (see, e.g., reviews [2], [5], [8], [13], [20], [22], [36], [46], [52]). These questions are interesting to practitioners as well as theoreticians, because the small-sample effects can easily contaminate the design and evaluation of a proposed system. For applications with a large number of features and a complex classification rule, the training sample size must be quite large. A large test sample is particularly essential to accurately evaluate a classifier with a very low error rate.

The design of a pattern recognition system consists of several stages: data collection, formation of the pattern classes, feature selection, specification of the classification algorithm, and estimation of the classification error. In this paper, we will discuss the effects of sample size on feature selection and error estimation for several types of classifier. In addition to surveying prior work in this area, our emphasis is on giving practical advice to today's designers and users of statistical pattern recognition systems. The paper is organized as follows. Section II introduces the classifiers we will focus on in this paper. In Section III, we explore classifier design in the context of small design sample size. The estimation of error rates under small test sample size follows in Section IV. Section V investigates sample size effects in feature selection. Section VI presents recommendations for the choice of learning and test sample sizes. Section VII contains some recommendations for the designers of pattern recognition systems. We make use of the following notation in this paper.

- $X = [X_1, \cdots, X_p]^T$: a *pattern* (or *feature vector*). The individual *features* are the X_i.
- p: the *dimensionality* (number of features).
- q_i: the *prior probability* of class π_i.
- f_i: the *class-conditional density function* of class π_i.
- $g(X) = q_1 f_1(X) - q_2 f_2(X)$: the Bayes discriminant function.
- $\hat{g}(X) = q_1 \hat{f}_1(X) - q_2 \hat{f}_2(X)$: the sample-based discriminant function.
- $\hat{g}^{\alpha}(X)$: the sample-based discriminant function of a classifier α.
- PMC: The *probability of misclassification* (or *error rate*).
- $P_B = \int_{\Omega} \min \{q_1 f_1(X), q_2 f_2(X)\}\, d(X)$: Bayes PMC.

II. Classification Algorithms

In the pattern recognition literature, there are a large number of ways to use sample observations to design a classification rule. One can use a statistical decision function approach with the Gaussian or exponential family of distributions along with a dozen of structural forms for the covariance matrices. One can further assume the covariance matrices to be equal or different for the various pattern classes. Classical maximum likelihood approaches can be used to estimate the parameters of the probability density functions corresponding to the pattern classes. In case of complex multimodal pattern classes, one can use a number of modifications of piecewise linear, piecewise quadratic classification rules, artificial neural networks, nonparametric Parzen window classifiers, or K-NN classifiers. The latter two can differ in the metric used to define the distance between two pattern vectors, and in methods used to edit the learning sample. There are several versions of classifiers based on potential functions [9] differing in a family of transformations of the pattern vector X, and in the optimization criteria. There are at least eight types of pattern error functions used to evaluate an empirical risk function in order to design linear and piecewise linear discriminant functions and artificial neural networks. Nearly two dozen methods exist for finding classification rules using heuristics when different similarity measures are applied to define the similarity between vector X, and the class π_i. A number of statistical models and expansions are known for approximating discrete distributions which can be used to design the classification rule for observations characterized by discrete or mixed variables. Therefore, the total number of classification methods which have been proposed in the pattern recognition literature exceeds two hundred.

Below, we describe several important classifiers, which have seen practical use. We will concentrate on the two-class problem

Manuscript received August 31, 1989; revised August 25, 1990. Recommended for acceptance by S. L. Tanimoto. This work was supported in part by the National Science Foundation under Grant CDA-8806599.

S. J. Raudys is with the Institute of Mathematics and Cybernetics, Lithuanian Academy of Sciences, Vilnius 232600, USSR.

A. K. Jain is with the Department of Computer Science, Michigan State University, East Lansing, MI 48824.

IEEE Log Number 9040949.

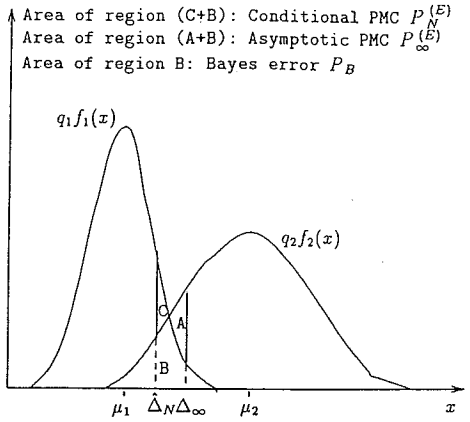


Fig. 1. Bayes PMC, asymptotic PMC, and conditional PMC for the Euclidean distance classifier. μ_1 and μ_2 are the true means, $\Delta_\infty = \frac{\mu_1+\mu_2}{2}$, and $\hat{\Delta}_N = \frac{\overline{x}^{(1)}+\overline{x}^{(2)}}{2}$, where $\overline{x}^{(1)}$ is the sample mean of class π_i; $\hat{\Delta}_N$ is a random variable.

in this paper. An unclassified p-dimensional multivariate feature vector $\boldsymbol{X}$ is allocated to the class π_1 if the discriminant function (DF) $g(\boldsymbol{X})$ is positive and to the class π_2 otherwise.

The quality of a classification rule will be characterized by its probability of misclassification (PMC). There are several definitions of PMC which are important in classifier design. In the following definitions of PMC, we are assuming that the number of test samples is infinite. In Section IV, we discuss the estimation of PMC when only a finite number of test samples is available.

- *Bayes PMC:* P_B is the PMC of an optimal Bayes classifier.
- *Conditional PMC:* P_N^α is the PMC of the classifier α trained on a given training sample of size N. The conditional PMC P_N^α is itself a random variable, since it is a function of the training samples. We assume that the training samples are labeled (i.e., we are working in the supervised learning mode).
- *Expected PMC:* EP_N^α is the expectation of P_N over all random training samples of sizes N_1 and N_2. $N = N_1 + N_2$.
- *Asymptotic PMC:* the probability of misclassification under the classifier α designed with an infinite number of training samples.

$$P_\infty^\alpha = \lim_{N_1, N_2 \to \infty} EP_N^\alpha \tag{1}$$

Note that P_∞^α can be different for different classifiers α. Also, $P_\infty^\alpha \geq P_B$ for all classifiers α.

Fig. 1 shows the Bayes error for a two-class one-dimensional problem, where q_i and $f_i(x)$ are the prior probability and class-conditional density of class π_i, $i = 1, 2$. The conditional PMC and the asymptotic PMC of the Euclidean distance classifier (described below) are also shown in Fig. 1.

We now briefly review six commonly used classifiers and provide the corresponding discriminant functions.

A. Euclidean Distance Classifier

This classifier makes classifications only according to sample means, $\overline{\boldsymbol{X}}^{(1)}$ and $\overline{\boldsymbol{X}}^{(2)}$ of the two classes. Its discriminant function is written as

$$\begin{aligned}\hat{g}^E(\boldsymbol{X}) &= \left(\boldsymbol{X} - \overline{\boldsymbol{X}}^{(2)}\right)^T \left(\boldsymbol{X} - \overline{\boldsymbol{X}}^{(2)}\right) \\ &\quad - \left(\boldsymbol{X} - \overline{\boldsymbol{X}}^{(1)}\right)^T \left(\boldsymbol{X} - \overline{\boldsymbol{X}}^{(1)}\right) \\ &= 2 \cdot \left[\boldsymbol{X} - \frac{1}{2}\left(\overline{\boldsymbol{X}}^{(1)} + \overline{\boldsymbol{X}}^{(2)}\right)\right]^T \left(\overline{\boldsymbol{X}}^{(1)} - \overline{\boldsymbol{X}}^{(2)}\right).\end{aligned} \tag{2}$$

The Euclidean distance classifier can be used when the pattern classes are well separated or when we want to implement a simple decision rule.

B. Fisher's Linear Discriminant

This is perhaps the most commonly used classification rule. The discriminant function is given by

$$\begin{aligned}\hat{g}^F(\boldsymbol{X}) &= \left[\boldsymbol{X} - \frac{1}{2}\left(\overline{\boldsymbol{X}}^{(1)} + \overline{\boldsymbol{X}}^{(2)}\right)\right]^T \boldsymbol{S}^{-1}\left(\overline{\boldsymbol{X}}^{(1)} - \overline{\boldsymbol{X}}^{(2)}\right) \\ &\quad + \ln \frac{q_1}{q_2},\end{aligned} \tag{3}$$

where q_1 and q_2 are the prior probabilities of the classes π_1 and π_2, respectively, and $\boldsymbol{S}$ is the sample covariance matrix (assumed to be common to both classes). It is an asymptotically optimal rule for the classification of Gaussian populations with a common covariance matrix.

C. Quadratic Discriminant Function

$$\begin{aligned}\hat{g}^Q(\boldsymbol{X}) &= \left(\boldsymbol{X} - \overline{\boldsymbol{X}}^{(2)}\right)^T \boldsymbol{S}_2^{-1}\left(\boldsymbol{X} - \overline{\boldsymbol{X}}^{(2)}\right) \\ &\quad - \left(\boldsymbol{X} - \overline{\boldsymbol{X}}^{(1)}\right)^T \boldsymbol{S}_1^{-1}\left(\boldsymbol{X} - \overline{\boldsymbol{X}}^{(1)}\right) + \ln \frac{|\boldsymbol{S}_2| q_1}{|\boldsymbol{S}_1| q_2},\end{aligned} \tag{4}$$

where $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are the sample estimates of the class-conditional covariance matrices, and $|\boldsymbol{S}_i|$ denotes the determinant of $\boldsymbol{S}_i$.

Many authors have noted that the linear discriminant function is robust to nonnormality of patterns [25], [30], [40], [42]. The quadratic DF, however, often significantly suffers from nonnormality of the data. There have been some attempts to obtain more robust discriminant functions (see, for example, the review in [6]). The quadratic DF (4) is a plug-in rule obtained from an optimal quadratic DF for two Gaussian populations where sample means and sample covariances have replaced the true parameters. But the resulting sample-based DF (4) is not "optimal" in the Bayesian sense [9]. It is important to note that, when training sample sizes N_1 and N_2 from the two classes are unequal, then the performance of the plug-in discriminant functions is further degraded [7]. For example, when the class-conditional distributions are multivariate normal with 40 features and true covariance matrices $\sum_2 = 2\sum_1$ and the asymptotic PMC is $P_\infty = 0.034$, then for training samples of sizes $N_1 = N_2 = 200$, the expected PMC $EP_N^Q = 0.151$. However, when $N_1 = 200$ and $N_2 = 2000$, $EP_N^Q = 0.169$ [18]. The main reason for the nonoptimality of the plug-in discriminant function is its bias [7], [22]. The use of unbiased estimates of multivariate Gaussian densities [1] or Bayes density estimates [27], [15] also result in biased discriminant functions and does not improve the performance of the quadratic discriminant function when $N_1 \neq N_2$. Grabauskas [18] found the expected value of (4) and

proposed an unbiased quadratic discriminant function, which has the form

$$\hat{g}^{QU}(\boldsymbol{X}) = \sum_{j=1}^{2}(-1)^{j+1}\left\{\left(1-\frac{p}{N_j}\right)\left(\boldsymbol{X}-\overline{\boldsymbol{X}}^{(j)}\right)^T \boldsymbol{S}_j^{-1} \left(\boldsymbol{X}-\overline{\boldsymbol{X}}^{(j)}\right) + \ln\frac{|\boldsymbol{S}_j|}{q_j} - \sum_{i=1}^{p}\Psi\left(\frac{N_j-i}{2}\right) + p\ln N_j\right\}, \tag{5}$$

where $\Psi(r)$ is the Euler Ψ function:

$$\Psi(r+1) = -C + \sum_{s=1}^{r}\frac{1}{s}$$
$$\Psi\left(r+\frac{1}{2}\right) = -C - 2\ln 2 + 2\sum_{s=1}^{r}\frac{1}{1+2s}. \tag{6}$$

This discriminant function has a lower error rate than $\hat{g}^{Q}(\boldsymbol{X})$ in the finite training sample case: $EP_N^{QU} = 0.096$ when $N_2 = N_1 = 200$, and $EP_N^{QU} = 0.085$ when $N_1 = 200$ and $N_2 = 2000$ ($\sum_2 = 2\sum_1$, $p = 40$, $P_\infty^Q = 0.034$).

D. Parzen Window Classifier

The Parzen window classifier does not assume a particular form for the class-conditional densities. Its discriminant function is

$$\hat{g}^{P}(\boldsymbol{X}) = \frac{q_1}{N_1}\sum_{i=1}^{N_1}K\left(\frac{\boldsymbol{X}-\boldsymbol{X}_i^{(1)}}{\lambda}\right) - \frac{q_2}{N_2}\sum_{i=1}^{N_2}K\left(\frac{\boldsymbol{X}-\boldsymbol{X}_i^{(2)}}{\lambda}\right), \tag{7}$$

and depends on the window function $K(\cdot)$ and on the value of the smoothing parameter λ. The most popular window functions are the exponential window

$$K_e\left(\frac{\boldsymbol{X}-\boldsymbol{Y}}{\lambda}\right) = \exp\left(-\frac{(\boldsymbol{X}-\boldsymbol{Y})^T(\boldsymbol{X}-\boldsymbol{Y})}{\lambda^2}\right) \tag{8}$$

and the logistic window

$$K_l\left(\frac{\boldsymbol{X}-\boldsymbol{Y}}{\lambda}\right) = \frac{\lambda^2}{\lambda^2+(\boldsymbol{X}-\boldsymbol{Y})^T(\boldsymbol{X}-\boldsymbol{Y})}. \tag{9}$$

Skurikhina [53] has tested 13 types of window functions (including Gaussian, logistic, trapezoidal, triangular, and sinusoidal) and found that with the proper selection of smoothing parameters, all 13 classifiers had nearly equal error rates. However, the value of the smoothing parameter is very important. It has been proved theoretically [9], [57] that the value of λ should decrease with an increase in the design sample size N. The Taylor series expansion of the window function in (8) yields

$$\hat{g}^{P}(\boldsymbol{X}) \to \hat{g}^{E}(\boldsymbol{X}) + \mathrm{tr}(\boldsymbol{S}_2) - \mathrm{tr}(\boldsymbol{S}_1), \tag{10}$$

when $\lambda \to \infty$, where $\mathrm{tr}(\boldsymbol{S})$ denotes the trace of the matrix $\boldsymbol{S}$. Thus, the Parzen window classifier becomes similar to the Euclidean distance classifier as the window width increases. On the other hand, when $\lambda \to 0$ the Parzen window classifier with the exponential or logistic window coincides with the 1-NN classification rule. An optimal value of λ which minimizes classification error depends both on the design sample size and on the distribution of the pattern vectors $f_i(\boldsymbol{X})$. When we have two Gaussian populations with equal covariance matrices, the optimal decision boundary is a hyperplane and λ should be very large. When we have complex multimodal distributions and the decision boundary is extremely nonlinear, then even for small design sample sizes we have to use a small value of the parameter λ. The shape of the curve showing the dependence of the expected probability of misclassification on the value of the smoothing parameter λ, depends significantly on the true probability functions of the classes (Configuration of an optimal Bayes decision boundary). Several typical curves are presented in Fig. 2. They show the importance of the problem of determination of optimal value of the smoothing parameter λ_{opt}. Many different criteria have been used to find λ_{opt} [23]. Theoretical considerations can show only the qualitative characteristics of the dependence of the optimal value of the smoothing parameter on dimensionality and sample size. It is impossible to find an optimal $\lambda, \lambda_{\mathrm{opt}}$, which minimizes the error rate for all class-conditional densities. In order to find λ_{opt} for a particular problem, we recommend evaluating the classifier's performance for several values of λ and choosing that value which provides the best performance. When the variances of all the p features in $\boldsymbol{X}$ differ significantly and we use the same λ for all the features then it is often necessary to normalize the features prior to use in the classifier. In such a case, λ_{opt} often falls in the interval $(0.01, 10.0)$. In most practical problems, the valley of the dependence curve $EP_N = \phi_N(\lambda)$ is rather flat. Thus, a set of ten values: $0.001, 0.01, 0.03, 0.1, 0.3, 1, 3, 10, 100, 1000$ is often sufficient to empirically determine λ_{opt}.

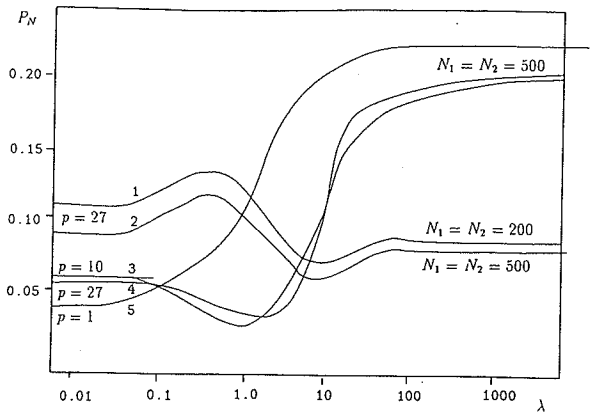


Fig. 2. Dependence of the conditional PMC (P_N) on the value of the smoothing parameter λ (data obtained from the 1976 Pattern Recognition Competition [3]: curves 1 and 2 correspond to original 27-variate data; curves 3, 4, and 5 correspond to the situation when the same data has been projected along its principal components).

Evaluation of the classification errors of Parzen window classifiers for ten different values of λ can be computationally demanding. Most of the computation time is spent in calculating distances $d\left(\boldsymbol{X}, \boldsymbol{X}_j^{(i)}\right)$ between the input pattern $\boldsymbol{X}$ from the test sample and $\boldsymbol{X}_j^{(i)}$, the jth training pattern from the ith class.

$$d\left(\boldsymbol{X}, \boldsymbol{X}_j^{(i)}\right) = \left(\boldsymbol{X}-\boldsymbol{X}_j^{(i)}\right)^T\left(\boldsymbol{X}-\boldsymbol{X}_j^{(i)}\right) = \sum_{s=1}^{p}\left(X_s - X_{js}^{(i)}\right)^2 \tag{11}$$

To conserve computer time, we recommend that all ten error rates be estimated simultaneously. After finding the dis-

tance $d\left(\boldsymbol{X}, \boldsymbol{X}_j^{(i)}\right)$, we calculate the ten terms $\frac{d\left(\boldsymbol{X}, \boldsymbol{X}_j^{(i)}\right)}{\lambda_r}$, $r = 1, \cdots, 10$ corresponding to each value of λ_r and classify the vector $\boldsymbol{X}$ for each value of λ_r. The computer time required to obtain ten simultaneous estimates of error rates is considerably smaller than the time for ten independently obtained estimates.

E. K-Nearest-Neighbor (K-NN) Classifier

In the K-NN rule, the class of the input pattern $\boldsymbol{X}$ is chosen as the class of the majority of its K nearest neighbors. The Euclidean metric is commonly used for distance calculations; however, the Mahalanobis metric can sometimes lead to better performance. The K-NN and Parzen window classifiers have many similar characteristics. It can be said that the K-NN classifier is the Parzen window classifier with a hyper-rectangular window function in the p-dimensional feature space. Both classifiers allow us to obtain complex nonlinear decision boundaries. The curvature of the boundary depends on the value of the smoothing parameters K and λ. When $K = 1$ or $\lambda \to 0$, the curvature is maximum; it diminishes with the increase of K or λ. The performance of K-NN classifier in finite design sample case significantly depends on the number K of nearest neighbors. Analogous to the Parzen window classifier, we recommend estimating the classification error for several values of K simultaneously, to find an optimal value of K.

F. Multinomial Classifier

This classifier is used for the recognition of patterns described by discrete variables. Let the jth variable take m_j distinct values. The p-variate vector $\boldsymbol{X}$, can therefore take one of $m = m_1 m_2 \cdots m_p$ values (states). Let p_{ij}, $i = 1, 2$; $j = 1, \cdots, m$ be the probability that a pattern from the class π_i takes on the jth state. Then the optimal Bayes discriminant function is given by

$$g^M(\boldsymbol{X}) = q_1 p_1 \boldsymbol{X}^s - q_2 p_2 \boldsymbol{X}^s, \tag{12}$$

where $\boldsymbol{X}^s$ denotes the label of the state corresponding to $\boldsymbol{X}$. In practice, instead of the true probabilities p_{ij}, one uses sample estimates

$$\hat{p}_{ij} = \frac{n_{ij}}{N_i}, \tag{13}$$

where n_{ij} stands for the number of cases when the pattern vectors from the design sample of the class π_i have taken on the jth state, resulting in the sample-based multinomial discriminant function $\hat{g}^M(\boldsymbol{X})$.

Sometimes the multinomial classifier is applied to continuous variables after making them discrete. One example of such a classifier is the *histogram classifier,* where for each class we design a histogram containing m bins. Such a classifier is very similar to the Parzen window classifier with rectangular windows in *a priori* fixed positions. Discretization of variables causes a loss of information. Therefore, the classification error of such a classifier is greater than that of the optimal Bayesian classifier. Obviously, the classification error depends on the number of bins $m_1, m_2, \cdots, m_p$ and on the values of the thresholds used for discretization of variables. The interaction between these parameters has not yet been explored. In one of our simulation studies with artificial Gaussian data ($p = 6$; only two variables are correlated), classification error of the quadratic discriminant function was $P_N^Q = 0.018$, while after discretization ($m_1 = 6$, $m_2 = 4$, $m_3 = m_4 = m_5 = m_6 = 2$), the multinomial classifier M resulted in $P_N^M = 0.058$.

When the dimensionality p is not small, then even in the binary case ($m_i = 2$), the total number of states ($m = 2^p$) is very large and it is difficult to obtain reliable estimates p_{ij}. In such a case, one needs to introduce some additional information in order to simplify the design. One possible way to do this is to assume that the variables are independent. Another alternative is to reduce the number of states m by designing a decision tree classifier.

A decision tree consists of a root node, intermediate nodes and $m^*(m^* << m)$ terminal nodes. At the root node, the best feature performs the decision. At the intermediate nodes, different features may participate at the same level. The final classification of the discrete vector $\boldsymbol{X}$ is performed according to the class number attributed to each particular terminal node. There are several methods to construct decision trees (see, e.g., [5], [31], [55], [51]). An advantage of such a classification rule is its applicability for classification of objects described by mixed variables and a comparatively easier interpretation of the classification results.

III. Sensitivity of Classifiers to Design Sample Size

In the finite design sample case, the parameters of the classifiers are estimated with low accuracy. Therefore, the resulting plug-in classifiers differ from optimal ones, resulting in an increase in the classification error. An increase in the classification error due to the finiteness of the design sample size $\Delta_N^\alpha = EP_N^\alpha - P_\infty^\alpha$ depends, first of all, on the type of the classification rule α, on the number of features p, and further, on the value of the asymptotic probability of misclassification. Significant research efforts have been made to find the relationship between classification error, learning sample size, dimensionality and complexity of the classification algorithm (see, for example, the reviews in [2], [10], [17], [22], [26], [38], [46], [45], [52], [56]).

In Table I, we present the number of observations required from each of the two classes to ensure that

$$\frac{EP_N^\alpha}{P_\infty^\alpha} \leq 1.5, \tag{14}$$

that is, the classification error due to the finiteness of the design sample size increases on the average 50% or less. These estimates of sample size N are calculated for the simple case of two Gaussian populations $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{I})$, with common identity covariance matrix and $N_1 = N_2 = \frac{N}{2}$. Note that when $\sum_1 = \sum_2 = \boldsymbol{I}$, only the Mahalanobis distance $\delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ is important in determining the increase in expected classification error EP_N for classifiers E, F, Q, and P. Therefore, the means $\boldsymbol{\mu}_i$ were chosen in such a way that the asymptotic probability of misclassification

$$P_B = P_\infty = \phi\left\{-\frac{1}{2}\sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}\right\} = 0.1 \quad \text{or} \quad 0.01. \tag{15}$$

For the parametric classifiers, these minimum values of N were obtained from analytical investigations [45]. For the Parzen window classifier, we used simulation studies, and for the multinomial classifier we obtained the required values of N analytically for the case of a "quasiuniform" distribution of probabilities p_{ij}, where p_{ij} can take only two values [19]:

$$p_1 = \frac{2P_\infty^{(M)}}{m} \quad \text{or}$$

$$p_2 = \frac{2\left(1 - P_\infty^{(M)}\right)}{m}.$$

TABLE I
DESIGN SAMPLE SIZE N REQUIRED FOR $\frac{EP_N^\alpha}{P_\infty^\alpha} \leq 1.5$. p IS THE DIMENSIONALITY.

(a) ASYMPTOTIC PMC, $P_\infty^{(\alpha)} = 0.1$. EXPECTED PMC, $EP_N^{(\alpha)} = 0.15$.
(b) ASYMPTOTIC PMC, $P_\infty = 0.01$. EXPECTED PMC, $EP_N = 0.015$.

Classifier	Sample Size ($N_1 = N_2 = \frac{N}{2}$)
E	$N = 1.2p$
F	$N = 4.0p$
Q	$N = \begin{cases} 80 = 10p, & \text{when } p = 8 \\ 320 = 16p, & \text{when } p = 20 \\ 1600 = 32p, & \text{when } p = 50 \end{cases}$
P	$N = 4.4(1.77)^p = \begin{cases} 24, & \text{when } p = 3, \lambda = 0.8 \\ 76, & \text{when } p = 5, \lambda = 0.8 \\ 420, & \text{when } p = 8, \lambda = 0.8 \end{cases}$ $N = 60$, when $p = 3$, $\lambda = 0.1$ $N > 100p$, when $p \geq 5$, $\lambda = 0.1$
M	$N = 3.3m$, when $10 \leq m \leq 100$

(a)

E	$N = 1.6p$
F	$N = 9.0p$
Q	$N = \begin{cases} 128 = 16p, & \text{when } p = 8 \\ 440 = 22p, & \text{when } p = 20 \\ 2000 = 40p, & \text{when } p = 50 \end{cases}$
P	$N = 30(1.55)^p = \begin{cases} 110, & \text{when } p = 3, \lambda = 0.8 \\ 280, & \text{when } p = 5, \lambda = 0.8 \\ 1000, & \text{when } p = 8, \lambda = 0.8 \end{cases}$ $N = 330$, when $p = 3$, $\lambda = 0.1$ $N >> 100p$, when $p \geq 5$, $\lambda = 0.1$
M	$N = 5.0m$, when $10 \leq m \leq 100$

(b)

While designing the parametric classifiers, each parameter estimate introduces its own contribution to the increase in the classification error. Below, we present an asymptotic formula for the increase in the expected PMC of parametric classifiers [46] under the assumption of Gaussian class conditional densities.

$$\Delta_N^\alpha = EP_N^\alpha - P_\infty^\alpha = \frac{1}{N}\frac{\varphi\left(\frac{\delta}{2}\right)}{\delta}\sum_{i \in C_\alpha} \theta_i, \tag{16}$$

where $\varphi(t) = (2\pi)^{-\frac{1}{2}} e^{\frac{-t^2}{2}}$. In Table II, δ^2 is the squared Mahalanobis distance between the class means μ_i, $P_\infty^F = \phi\left(-\frac{\delta}{2}\right)$, and $\theta_1, \theta_2, \cdots$ are the contribution terms resulting from the estimation of prior probabilities of the classes, q_1, q_2 (term θ_1), means of the classes, μ_1, μ_2 (term θ_2), covariance common for both classes (term θ_5), etc. The number of terms in set C_α in (16) depends on the classifier type α. Values of the terms presented in Table II show explicitly the increase in error due to the estimation of various parameters of the multivariate Gaussian distribution from the learning sample data. In (2), we used only sample means of the two classes. Therefore, for the Euclidean distance classifier, only the θ_2 term appears in (16). For the linear Fisher discriminant function (3), we use the estimates of the priors, means and common covariance matrix. Therefore, here we should use the terms θ_1, θ_2, and θ_5. Analogously, for the quadratic discriminant function (4), we should use the terms θ_1, θ_2, and θ_6.

Equation (16) shows that the increase in classification error of the parametric classifiers is proportional to $\frac{1}{N}$ and depends on the dimensionality of the feature vector p; for the linear classifiers, the relationship is linear and for the quadratic classifier the relationship is quadratic (only for large p when $p \gg \delta^2$). Analytical and simulation studies show that for the nonparametric classifiers (Parzen and multinomial), the decrease of $\Delta_N = EP_N - P_\infty$ with an increase in the design sample size N is slower $\left(O\left(\frac{1}{\sqrt{N}}\right)\right.$ or $\left.O\left(\frac{1}{N^{1/3}}\right)\right)$. For large values of the smoothing parameter λ (when the Parzen window classifier becomes similar to the Euclidean classifier), the decrease of Δ_N is of $O\left(\frac{1}{N}\right)$. Our theoretical and simulation studies have shown that when we use the same smoothing parameter λ for all features, then the increase in classification error Δ_N of the Parzen window classifier depends not on the actual dimensionality p, but on the intrinsic dimensionality p^* of the patterns [39]. The analysis also shows that the design sample size required to achieve a learning accuracy determined *a priori*, depends on the dimensionality exponentially:

$$N = \alpha\beta^{p^*}, \tag{17}$$

where scalars α and β depend on asymptotic and expected probabilities of misclassification, and on the value of the smoothing parameter (see Table I). The required design sample size for multinomial classifier depends linearly on the number of states m (when the distribution of the probabilities p_{ij} is "quasiuniform"):

$$N = m\gamma \tag{18}$$

where γ is a data-dependent constant.

Suppose that discrete values of the variables are obtained by discretization of each variable into r states. Then $m = r^p$ and, like in Parzen window classifier, we again have

$$N = \gamma r^p. \tag{19}$$

In real problems, many states have nearly zero probabilities p_{ij}. Sample estimates (13) of the probabilities of the rare states are not reliable but since they are seldom observed, rare states increase the classification error negligibly. The main increase in the classification error is caused by states with large probabilities. Therefore, it can be said, that increase in classification error depends not on the given number of states m, but on the effective number of states m^*. Usually $m^* \ll m$; however, exact knowledge of m^* and methods for its estimation are not known.

Estimates of the design sample sizes have been obtained for the case of Gaussian distributions with identical covariance matrices. In reality, additional factors effect the increase in the classification error, such as unequal covariance matrices and unequal design sample sizes from both populations. Therefore, the above estimates only provide some guidelines. Moreover, these relations between sample size and dimensionality are determined for a fixed value of asymptotic PMC. While solving real pattern recognition problems, P_∞ decreases with the addition of new variables, but then the problem of determining optimal number of features arises (see Section V).

An important quantity is the variance of the conditional probability of misclassification, $V(P_N)$. From Efron's analysis [10], it follows that for several parametric linear classifiers (Fisher's discriminant, logistic regression, and the Euclidean distance classifier), the increase in classification error $\Delta_N = P_N - P_\infty$ is distributed as a scaled chi-squared random variable $c\chi^2/N$ with p degrees of freedom, where the constant c depends on the asymptotic probability of misclassification and on the type of classification rule. Thus the ratio of the standard error of P_N to

TABLE II
CONTRIBUTION TERMS IN (16) NEEDED TO COMPENSATE FOR THE ESTIMATION OF VARIOUS PARAMETERS OF MULTIVARIATE GAUSSIAN DENSITIES.

i	θ_i	Parameters to Be Estimated from the Learning Sample
1	1	Prior probabilities q_1, q_2.
2	$\frac{\delta^2}{2}+p$	Means μ_1, μ_2.
3	$\frac{\delta^4}{8}+\frac{\delta^2}{2}$	Variances of the populations common for two classes (p parameters of a diagonal matrix).
4	$\frac{\delta^4}{4}+\frac{3\delta^2}{2}+p$	Variances of the populations different for two classes ($2p$ parameters of two diagonal matrices).
5	$\frac{\frac{\delta^4}{8}+\frac{p\delta^2}{4}}{\left(1-\frac{p}{N}\right)}$	Covariance matrix Σ, common to both classes ($\frac{p(p+1)}{2}$ parameters).
6	$\frac{2\left(\frac{\delta^4}{8}+\frac{p\left(p+\delta^2\right)}{4}\right)}{\left(1-\frac{2p}{N}\right)}$	Covariance matrices Σ_1 and Σ_2 different for both classes ($p(p+1)$ parameters.)

the mean increase in PMC $\sqrt{V(P_N)}/(EP_N - P_\infty) = \sqrt{(2/p)}$, which tends to zero as dimensionality increases.

IV. PERFORMANCE ESTIMATION

A number of methods for estimating the classification error have been proposed in the literature reviews [13], [16], [20], [29], [33], [34], [36], [43], [47], [54]. These methods can be studied by using the following two factors:

- The way in which multivariate observations are used to design the classifier and to test its performance;
- The pattern error function that determines the contribution of each observation of the test set to the estimate of the probability of misclassification.

There are four main approaches to use the given observations as the design set and as the test set.

1) The *Resubstitution Method* $\mathcal{R}$: all observations are used to design the classifier and used again to estimate its performance.
2) The *Hold-Out Method* $\mathcal{H}$: Suppose the total number of available observations is n^*. One portion of the set of observations (the *design set* containing N observations) is used to design the classifier, and the remaining $(n^* - N)$ portion (the *test set*) is used to estimate the error rate.
3) The *Cross-Validation Method* $\mathcal{L}$: In this method, $\binom{n^*}{k}$ classifiers are designed. Each classifier is designed by choosing k of the n^* observations as a design set, and its error rate is estimated using the remaining $(n^* - k)$ observations. This process is repeated for all distinct choices of k patterns and the average of the error rates is computed. A popular choice for the value of k is $k = 1$, yielding the well-known leave-one-out method.
4) The *Bootstrap Method* $\mathcal{B}$: A *bootstrap design sample* of size N is formed from the N observations by sampling with replacement. The classification rule is designed using this bootstrap sample and is tested twice:
   - N observations of the bootstrap design sample are used to obtain a bootstrap resubstitution estimate $P_R^{\mathcal{B}}$; and
   - the original design set is used to obtain the bootstrap estimate of conditional error $P_N^{\mathcal{B}}$.

   This procedure is repeated r times (typically, r lies between 10 and 200). An arithmetic mean $\overline{\Delta}_{N\mathcal{R}}^{\mathcal{B}}$ of the differences

$$\Delta_{N\mathcal{R}}^{\mathcal{B}^i} = P_N^{\mathcal{B}^i} - P_{\mathcal{R}}^{\mathcal{B}^i}, \qquad i = 1, \cdots, r \tag{20}$$

   is used to reduce the optimistic bias of the resubstitution estimate:

$$\hat{P}_{\mathcal{B}} = \hat{P}_{\mathcal{R}} + \overline{\Delta}_{N\mathcal{R}}^{\mathcal{B}}. \tag{21}$$

There are many modifications of the bootstrap method: the randomized bootstrap, the 0.632 estimator, the MC estimator, the complex bootstrap (see, e.g., [20], [21]).

Each of the above error estimation procedures can be used with different pattern error functions $h(\hat{g}(\boldsymbol{X}))$, where $\hat{g}(\boldsymbol{X})$ is the sample-based discriminant function:

$$\hat{P} = \frac{1}{n_t}\sum_{j=1}^{n_t} h(\hat{g}(\boldsymbol{X}_j)), \tag{22}$$

and $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_{n_t}$ are test sample observations.

1) *Error Counting* (EC):

$$h^{EC}(\hat{g}(\boldsymbol{X})) = \begin{cases} 1 & \text{if } \hat{g}(\boldsymbol{X}) < 0 \text{ and } \boldsymbol{X} \in \pi_1 \\ 1 & \text{if } \hat{g}(\boldsymbol{X}) > 0 \text{ and } \boldsymbol{X} \in \pi_2 \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

   Here, correctly recognized observations do not affect the estimate of PMC.

2) *Smooth Modification of EC* (SM) [16]:

$$h^{SM}(\hat{g}(\boldsymbol{X})) = \begin{cases} 1 & \text{if } \hat{g}(\boldsymbol{X}) \geq a \text{ and } \boldsymbol{X} \in \pi_2 \\ 1 & \text{if } \hat{g}(\boldsymbol{X}) \leq -a \text{ and } \boldsymbol{X} \in \pi_1 \\ 1 - \frac{\hat{g}(\boldsymbol{X})+a}{b} & \text{if } -a < \hat{g}(\boldsymbol{X}) < b - a \text{ and } \boldsymbol{X} \in \pi_1 \\ 1 + \frac{\hat{g}(\boldsymbol{X})-a}{b} & \text{if } a - b < \hat{g}(\boldsymbol{X}) < a \text{ and } \boldsymbol{X} \in \pi_2 \\ 0 & \text{otherwise.} \end{cases} \tag{24}$$

Here, part of the correctly classified observations contribute to the estimation of misclassification probability.

3) *Posterior probability estimate* (PP) [14], [32]:

$$h^{PP}(\hat{g}(\boldsymbol{X})) = \frac{1}{2}[1 - |\tanh(\hat{g}(\boldsymbol{X})/2)|], \tag{25}$$

where the $\hat{g}(\boldsymbol{X})$ defined in (25) is given by

$$\hat{g}(\boldsymbol{X}) = \frac{q_1 \hat{f}_1(\boldsymbol{X})}{q_2 \hat{f}_2(\boldsymbol{X})} \tag{26}$$

and $\hat{f}_i(\boldsymbol{X})$ is a sample estimate of the probability density function $f_i(\boldsymbol{X})$. An advantage of this estimate is that the test sample observations can be unlabeled. Information about the design sample propagates into the error estimate as well.

4) *Quasiparametric estimate* (QP) [29]: Here it is assumed that the values of the discriminant function $\hat{g}(\boldsymbol{X})$ have a Gaussian distribution. PMC is found analytically from sample means and variances of the values $\hat{g}(\boldsymbol{X}_j^{(i)})$, $i = 1, 2$ and $j = 1, \cdots, N_i$.

Thus, in principle, we can have 16 error estimation methods. Additionally, there is a series of parametric methods generally applied to cases where one can assume a parametrized family of distributions for the class conditional densities, usually Gaussian with common covariance matrices [33], [50]. Which method should one use? In spite of numerous research efforts, only weak recommendations can be given to practitioners.

The resubstitution method results in optimistically biased estimates of the asymptotic PMC P_∞. Therefore, it can be used only when the sample size is sufficiently large. It was shown analytically [12], [43] (for Euclidean and Fisher classifiers) and experimentally that the bias of the resubstitution estimate ($\Delta_{\mathcal{R}} = P_\infty - E\hat{P}_{\mathcal{R}}$) is approximately equal to the bias of the expected PMC ($\Delta_N = EP_N - P_\infty$).

A symmetry property of the expectation of resubstitution estimate $E\hat{P}_{\mathcal{R}}$ and expected PMC EP_N (see Fig. 3) shows that the estimates of sufficiency of the design sample size (Table I) can be used to determine conditions when the resubstitution method can be used.

The hold-out error counting estimate results in an unbiased estimate of the expected PMC. The disadvantage of this method is that not all observations of the design sample take part in the learning process and only a part of observations are used to evaluate the classification error.

The leave-one-out error counting procedure produces a practically unbiased estimate of the expected PMC if the sample observations are statistically independent. In case of dependent observations this method approaches the resubstitution method and results in an optimistically biased estimate of the expected PMC, EP_N. The leave-one-out estimate $\hat{P}_{\mathcal{L}}$ can be used together with resubstitution estimate $\hat{P}_{\mathcal{R}}$ in order to get an estimate of the asymptotic PMC:

$$\hat{P}_\infty = \frac{\hat{P}_{\mathcal{L}} + \hat{P}_{\mathcal{R}}}{2}. \tag{27}$$

The above estimate follows from the symmetry of the curves $E\hat{P}_{\mathcal{R}} = \phi_1(N)$ and $EP_N = E\hat{P}_{\mathcal{L}} = \phi_2(N)$ in Fig. 3.

A disadvantage of the leave-one-out method is that for some types of classification algorithms, it requires substantial computation time compared to the hold-out and resubstitution methods. For many algorithms (e.g., all algorithms described in Section II),

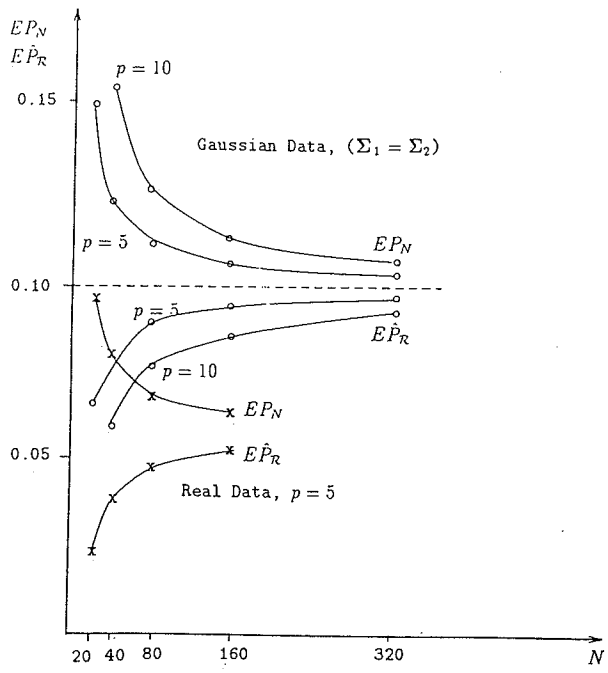


Fig. 3. Dependence of the expected PMC and the expectation of resubstitution estimate on the sample size for a linear discriminant function. Error estimates for Gaussian data are shown by circles, and for a real data set are shown by crosses.

however, special recursive relations allow the classifier to be designed N times with negligible increase in computation. Bootstrap methods and their variants appear more accurate than the leave-one-out method only when the classification error is large [35], [47], [48].

Analytical and experimental investigations of the error-counting methods show that the variance is of order [43], [48], [47]

$$V\hat{P}_\eta = \frac{E\hat{P}_\eta\left(1 - E\hat{P}_\eta\right)}{n_t}, \tag{28}$$

where $E\hat{P}_\eta$ and $V\hat{P}_\eta$ are the mean and variance of the error estimate $\hat{P}_\eta$, and η indicates the method: $\mathcal{R}$, $\mathcal{H}$, $\mathcal{L}$, or $\mathcal{B}$; n_t is the number of test samples (for $\mathcal{R}$, $\mathcal{L}$, and $\mathcal{B}$ methods, $n_t = N$). Since the resubstitution method is optimistically biased, $E\hat{P}_{\mathcal{R}}$ takes the smallest value. Therefore, according to (28), the resubstitution method also has the smallest variance.

The variance of the SM, PP, and QP estimates can be less than the variance of the error-counting estimate. However, the SM, PP, and QP estimates are often biased. The bias of the SM estimate depends directly on the degree of smoothing of the pattern error function [parameters a and b in (24)] and can exceed the absolute value of the classification error. It was observed experimentally [43] that the QP estimate is pessimistically biased in the low-dimensional case, when the distribution function of the discriminant function $\hat{g}(\boldsymbol{X})$ is non-Gaussian. The PP estimate is based on an information contained in the test sample and on additional information used to obtain estimates $\hat{f}_i(\boldsymbol{X})$ of the probability densities $f_i(\boldsymbol{X})$ of the pattern classes. In the parametric case, this information can be useful and can reduce the variance of the PP estimate. When the design sample size is small or when the additional information is incorrect (e.g., we assume normality for the class-conditional densities when in fact they are significantly non-Gaussian), then the estimated class-

conditional probability densities and consequently the pattern error function (25) are determined with large errors. This leads to significant bias of the PP estimates. The bias is especially large in the nonparametric case, where very vague prior information is typically used to obtain estimates of $\hat{f}_1(\boldsymbol{X})$ and $\hat{f}_2(\boldsymbol{X})$ in (25).

In feature selection (see Section V), the bias of the estimates of the classification error is not critical if it is approximately equal for all subsets of variables. However, in estimating the performance of a complete pattern recognition system, the use of the biased estimates is dangerous.

V. Feature Selection

One of the fundamental problems in statistical pattern recognition is to determine which features should be employed for the best classification results. The purpose of feature selection and extraction is to identify those features which are important in discriminating among pattern classes. The need to retain only a small number of *useful* and *good* features in designing a classifier has been well documented in the literature [9]. In this section, we describe the counterintuitive phenomenon of peaking in classification performance when the number of features is increased, and the classification error is used as a feature selection criterion.

A. Optimal Number of Features

In Section III, we examined the relationship between sample size and dimensionality for fixed asymptotic PMC, P_∞. While solving real pattern recognition problems, the addition of new features usually decreases P_∞. Usually, the "best" features are added first, and less-useful features are added later. Therefore, the rate of decrease in P_∞ slows as the number of features increases. Adding new features requires that new parameters be estimated. An inexact estimation of parameters increases classification error. If this increase is larger than the decrease in classification error produced by the addition of the new feature, then the net effect is that addition of the new feature increases the error rate. Therefore, in the finite-sample-size case we have a "peaking" phenomenon: classification error initially drops with addition of new features, then attains a minimum, and then begins to increase. The number of features at which the expected PMC, EP_N, is minimal is called the *optimal* number of features, and is denoted p_{opt}. It depends on the design sample size, the type of classification rule, the class-conditional distributions of the pattern vector $\boldsymbol{X}$, and most importantly, on the effectiveness of features and their ordering [26], [41]. In practice, it is important to know if the optimal number of features p_{opt} is lower than the initial number of features p. Typically, p_{opt} is smaller for smaller design sample sizes, for more complex classification algorithms, and for better orderings of features. When all features are equally effective, or when the features are unordered and added in a random way, p_{opt} can be large (for the linear discriminant function, $p_{\text{opt}} = \frac{N}{2} - 1$ for N training patterns [24]; for the quadratic discriminant function, p_{opt} is significantly lower than $\frac{N}{2} - 1$, but still increases with the number of training patterns N). When features or sets of features are ordered *a priori*, the following simple procedure can be used to estimate p_{opt}. Obtain *unbiased* estimates $\hat{P}_{(i)}$ of the expected PMC for $i = 1, \cdots, p$ and select that subset which produces the lowest estimate. Here, one can use the hold-out, leave-one-out, or even parametric unbiased estimators of the expected PMC [2], [11], [29], [33], [44], [56]. For example, for the Fisher linear discriminant, the following estimate of the expected PMC is used:

$$E\hat{P}_N^{(F)} = \Phi\left\{-\frac{\hat{\delta}}{2}\left[\left(1+\frac{p}{N_1+N_2-p}\right)\left(1+\frac{(p-1)(N_1+N_2)}{N_1N_2\hat{\delta}^2}\right)\right]^{-\frac{1}{2}}\right\}, \tag{29}$$

where $\hat{\delta}^2$ is an unbiased estimate of the Mahalanobis distance:

$$\hat{\delta}^2 = \left(\overline{\boldsymbol{X}}^{(1)} - \overline{\boldsymbol{X}}^{(2)}\right)^T \boldsymbol{S}^{-1}\left(\overline{\boldsymbol{X}}^{(1)} - \overline{\boldsymbol{X}}^{(2)}\right) \frac{N_1+N_2-p-3}{N_1+N_2-2} - \frac{N_1+N_2}{N_1N_2}p. \tag{30}$$

If the dependence $\hat{P}_{(i)} = \phi(i)$ is not smooth, then better results for p_{opt} can be obtained after smoothing the empirical dependence. This can be done by hand, or formally through use of a special mathematical technique [44]. Often, the dependence $EP_N(i) = f(i)$ is flat near the minimum point $EP_N(p_{\text{opt}})$. Therefore, the accuracy of determining p_{opt} is not crucial; it is more important to pay attention to the existence of the peaking effect and make minor efforts to determine p_{opt}.

B. Accuracy of Feature Selection

Usually, features are not ranked according to their effectiveness in discrimination *a priori*. We use the sample information to compare the effectiveness of features and rank them. The sample estimates of the effectiveness are not exact. Therefore, only the best features can be ranked properly. The effectiveness of the worst features differs negligibly and the accuracy of sample estimates is not sufficient for exact ranking of features or the feature sets. The inaccuracy of the estimates of the feature effectiveness causes a bias in the estimates of the best subsets containing $i = 1, \cdots, (p-1)$ features. Therefore, the estimates of p_{opt} become biased also. The problem of estimating p_{opt} in the case of empirical ordering of features is unsolved.

The ordering of features is an important step in the design of a pattern recognition algorithm. It is well known that the ordering of features and the ordering of feature subsets are two different subjects. In general, the best subset of t features and the set of t individually best features are not identical ($t < p$). The only procedure which guarantees that the best subset is found, is a complete inspection of all subsets, which is computationally expensive. Therefore, many suboptimal procedures for feature subset selection have appeared in the literature [28]. None of these techniques guarantee that the best feature subset will be found, but they typically require much less computation than exhaustive search. Most procedures use either sequential addition of features, sequential deletion of features, or a combination of both approaches. Other techniques include random search (inspection of randomly selected subsets), directed random search, and branch-and-bound techniques.

Each feature selection procedure can be carried out by using every one of a number of feature effectiveness criteria or the classification error estimation methods mentioned in the previous section.

There are approximately two dozen parametric criteria of feature effectiveness known in the pattern recognition literature [4], [58]. Examples are the Mahalanobis, Bhattacharyya, Patrick-Fisher, and Matusita distances, divergence, mutual Shannon information, and entropy. Analytical expressions of these criteria are usually simpler than the Bayes error expressions. Some of them use additional information about the populations to be

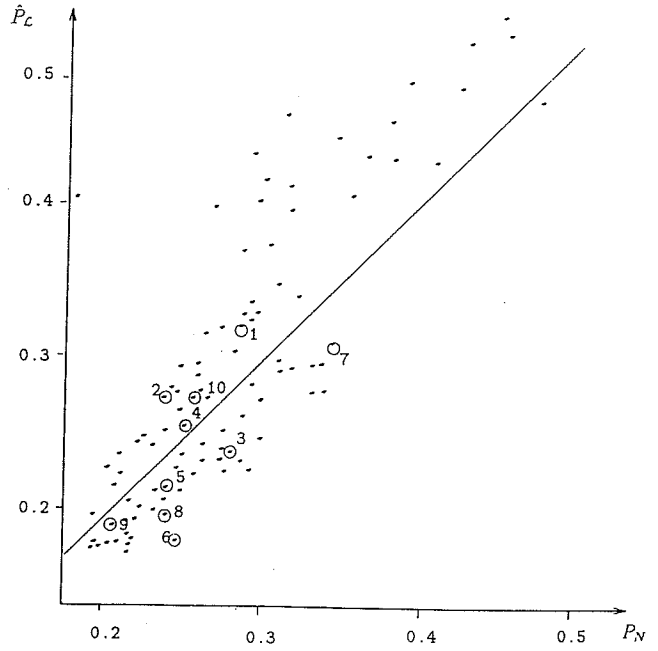


Fig. 4. Distribution of conditional PMC, P_N, and its *leave one out* error counting estimate $\hat{P}_L$ of 100 random 4-variate subsets of features (real medical data, 19 original features, Fisher's linear discriminant, $N_2 = N_1 = 25$. Conditional PMC P_N^F was estimated by holdout method using an additional 500 observations). The total number of feature subsets in this experiment is $m = 1000$.

classified. For example, the Mahalanobis distance criterion is based on assumptions of multivariate normality for all classes, with a common covariance matrix. The variances of criteria containing such assumptions are usually lower than variances of the nonparametric estimators described above; however, biased estimates are produced if the parametric assumptions are not valid. Therefore, the usefulness of the criteria with parametric assumptions can be justified only when the bias is approximately equal for all subsets of features. It is unclear whether one specific strategy provides consistently better performance than most others. One experimental comparison [37] concluded that forward selection and random search outperformed other procedures in one application.

The analysis of feature selection strategies is complicated by the fact that all feature effectiveness criteria are subject to error, caused by both sample size effects and the simplifying assumptions. Inaccurate criteria of effectiveness can lead to incorrect rankings of features and feature subsets. Therefore, one objective of feature selection is to find feature subsets which produce an expected PMC close to the ideal value (the value produced by the truly "optimal" feature subset).

In order to explain the mechanism of an increase in classification error due to nonoptimal feature selection, we shall at first explain three types of classification error occurring in the feature selection process.

Suppose we have m subsets of features $S_1, \cdots, S_m$ with corresponding effectiveness estimates $\hat{P}_1, \cdots, \hat{P}_m$. Suppose there exist true values of their effectiveness (e.g., the Bayes error, conditional PMC, expected PMC) $P_1, \cdots, P_m$. A 2-D scatter plot of the m ordered pairs $(\hat{P}_i, P_i)$ is depicted in Fig. 4. Let us first examine the ten random subsets of features denoted in Fig. 4 by circles. We can define three types of classification errors (note that the sense and the definition of these errors are different from that presented in the beginning of Section II or in Hand's [20] review):

1) *Apparent* error, $\hat{P}_{\text{apparent}}$, in selection, i.e., the minimum of $\hat{P}_1, \cdots, \hat{P}_m$. For the ten pairs in Fig. 4, we have

$$\hat{P}_{\text{apparent}} = \hat{P}_6 = 0.18. \tag{31}$$

2) *Ideal* PMC, P_{ideal}, the minimum of the true error probabilities $P_1, \cdots, P_m$. In the example,

$$P_{\text{ideal}} = P_9 = 0.220. \tag{32}$$

3) *True* PMC, P_{true}, the true error rate of that subset with the minimal error estimate $\hat{P}_i$. In the example,

$$P_{\text{true}} = P_6 = 0.242. \tag{33}$$

The "cloud" of points $(\hat{P}_i, P_i)$, $i = 1, \cdots, m$, shown in Fig. 4 also arises in cases where feature subsets are formed according to the exhaustive feature selection criterion, sequential feature addition and deletion, etc. The values P_{true}, P_{ideal}, and $\hat{P}_{\text{apparent}}$ are determined from the set of bivariate vectors $(\hat{P}_i, P_i)$, $i = 1, \cdots, m$. The set depends on the class-conditional densities $f_i(\boldsymbol{X})$, on the feature selection procedure used, and on the accuracy of the estimates $\hat{P}_i$. In Fig. 5 we show three curves which represent the dependence of the values P_{true}, P_{ideal}, and $\hat{P}_{\text{apparent}}$ on the number of subsets m for the data presented in Fig. 4. Subsets of variables were presented in a random way. For large m, we have $P_{\text{true}} > P_{\text{ideal}}$ and $\hat{P}_{\text{apparent}} < P_{\text{ideal}}$. In practice the value $\Delta_1 = P_{\text{true}} - P_{\text{ideal}}$ measures the "distance" between the feature selection procedure in use and the ideal procedure (it also allows P_{true} to be predicted from $\hat{P}_{\text{apparent}}$). This problem has only recently been investigated, and no strong conclusions have yet been drawn [49]. Preliminary theoretical and experimental investigations show that Δ_1 is of order

$$k\sqrt{\frac{P_{\min}(1 - P_{\min})}{n_t}}, \tag{34}$$

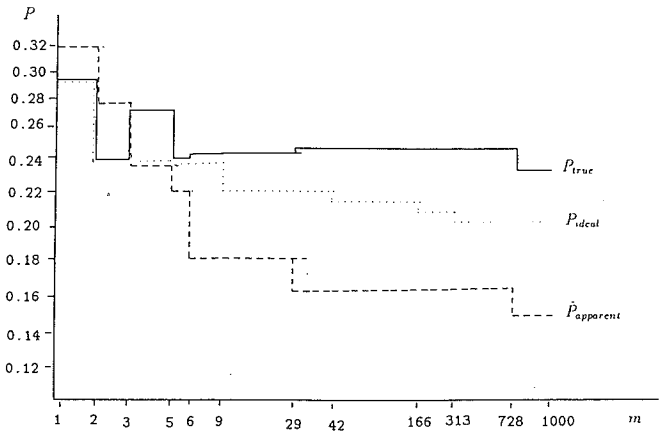


Fig. 5. Dependence of the true, ideal, and apparent errors on the number of subsets m.

and

$$P_{\text{true}} \approx \hat{P}_{\text{apparent}} + 2\Delta_1, \tag{35}$$

where k is a constant between 0.25 and 1, $P_{\min}$ denotes the PMC of the best subset of features, and n_t is the size of the test sample used to obtain $\hat{P}_1, \cdots, \hat{P}_m$. The values Δ_1 and $P_{\min}$ are random variables, with variance on the order of

$$\frac{P_{\min}(1 - P_{\min})}{n_t} \tag{36}$$

Therefore, Δ_1 and P_{true} can only be predicted with large error. It has been shown [36] that a four-fold increase in the test sample size n_t will halve Δ_1. The difference Δ_1 also increases with the number of feature subsets m in the random search feature selection procedure. The true classification error, P_{true}, decreases rapidly at first with an increase in m, but as m becomes large, the amount of decrease in P_{true} becomes small. Therefore, when the accuracy of the estimates $\hat{P}_1, \cdots, \hat{P}_m$ is low, there is no reason to increase the number of subsets m or to use a complex feature selection procedure. To improve the feature selection efficiency, one must use more accurate estimators of efficiency or obtain additional information about the class-conditional densities and incorporate this information in the estimates.

VI. Sample Size Determination

There are two occasions when the designer of a pattern recognition system has to determine the size of the sample:

1) to find a sample size sufficient to achieve a desired level of learning accuracy, and
2) to find the size of the test sample sufficient to estimate the classification error.

It was mentioned in Section III that the minimum design sample size depends on the method used to find coefficients of the classification rule, the number of features, the asymptotic PMC, and the desired learning accuracy. In Table I, we presented approximate sample sizes required to ensure that the expected PMC, EP_N, would not exceed the asymptotic PMC, P_∞, by more than $\beta = 50\%$. Equation (16) shows that in order to double the accuracy (i.e., $\beta = 25\%$), the design sample sizes should double for parametric classifiers and more than double for nonparametric classifiers. Estimates of the sufficient sample size depend slightly on the asymptotic PMC. Therefore, in order to use them in practice we have to estimate an interval in which P_∞ will lie. For example, we might guess that by using the Fisher's linear discriminant, $P_\infty \in [0.01, 0.1]$. Then, the requirement for the efficiency of the design sample requires $32 \leq N \leq 72$ for a dimensionality of eight. If we assume $P_\infty \in \{0.1, 0.5\}$, then $N \leq 32$ (see Table I).

The estimates of the minimum design sample size were obtained for some idealized (spherically Gaussian or *quasi-uniform*) class-conditional densities $f_i(\boldsymbol{X})$ (see Section III). For other distributions, the required design sample sizes can be different. Therefore, in order to estimate the sufficiency of the design sample size, we recommend additionally to use the following nonparametric estimate of the increase in the classification error $\Delta_N = EP_N - P_\infty$:

$$\hat{\Delta}_N = \frac{\hat{P}_{\mathcal{L}} - \hat{P}_{\mathcal{R}}}{2}, \tag{37}$$

where $\hat{P}_{\mathcal{L}}$ and $\hat{P}_{\mathcal{R}}$ are the leave-one-out and resubstitution estimates of the classification error, respectively. The estimate $\hat{\Delta}_N$ is based on the fact that the dependencies $EP_N = \phi_1(N)$ and $E\hat{P}_{\mathcal{R}} = \phi_2(N)$ are nearly symmetrical (see Section IV) and that $E\hat{P}_{\mathcal{L}} \approx EP_N$.

If the difference $\hat{\Delta}_N$ is small in comparison with the empirical estimate of the asymptotic PMC (27), then we can conclude that the design sample size is sufficient. Here, we have to pay attention to the variances of the estimates $\hat{\Delta}_N$ and $\hat{P}_\infty$. Extensive simulation studies have suggested that the estimates $\hat{P}_{\mathcal{L}}$ and $\hat{P}_{\mathcal{R}}$ are practically statistically independent. Therefore, the estimates of the variances and mean square errors (MSE) of $\hat{\Delta}_N$ and $\hat{P}_\infty$ can be found from (28):

$$MSE\left(\hat{\Delta}_N\right) = \frac{1}{2}\sqrt{\frac{\hat{P}_{\mathcal{L}}\left(1 - \hat{P}_{\mathcal{L}}\right)}{n_t} + \frac{\hat{P}_{\mathcal{R}}\left(1 - \hat{P}_{\mathcal{R}}\right)}{n_t}} = MSE\left(\hat{P}_\infty\right). \tag{38}$$

For example, when solving a pattern recognition problem with $N_1 = N_2 = 100$, (i.e., $N = 200$), with $\hat{P}_{\mathcal{L}} = 0.09$ and $\hat{P}_{\mathcal{R}} = 0.05$, then $\hat{P}_\infty = 0.07$, $\hat{\Delta}_N = 0.02$, $MSE(\hat{P}_\infty) = MSE(\hat{\Delta}_N) = 0.013$, i.e., the increase in the classification error is 30% of the asymptotic PMC. Therefore, we can conclude that the design sample size is sufficient. Note that in nonparametric estimation of the classification error, dimensionality does not play a role.

The size of the test sample n_t used to determine the performance of the classifier can be determined from the variance (28). If we require that the error counting estimate of the PMC does

not deviate from the true value P by more than $k\%$, then

$$2\sqrt{\frac{P(1-P)}{n_t}} = \frac{Pk}{100}. \tag{39}$$

Therefore, the estimate of the sufficient test sample size n_t is

$$n_t = \frac{4(1-P)}{P(k/100)^2}. \tag{40}$$

Here, we again have to guess at a possible value or interval of the true classification error. For example, if we assume $0.02 < P < 0.1$, then (40) produces an interval of $900 < n_t < 4900$ for $k = 20\%$.

While using the hold-out error estimation method we have to divide an existing set of observations into two parts: the design sample and the test sample. If the design sample is small, the classification error will be large. If the test sample is too small, then the variance of the error estimator will be large. In order to find an optimal balance between the sizes of the design and test samples, we have to introduce a loss function. One possible loss function is

$$\begin{aligned} LOSS(N_1, N_2) = {} & C_1(EP_N(N_1, N_2) - P_\infty) \\ & + C_2 MSE\left\{\hat{P}(n^* - N_1 - N_2)\right\}, \end{aligned} \tag{41}$$

where n^* is the total number of observations, $N_1 + N_2$ of which are used to design the classifier and the remainder used as the test sample, C_1 and C_2 are the costs associated with an increase in classification error (due to design sample size) and an increase in MSE of the error estimate (due to test sample size), respectively.

From the definition of the loss function above, it follows that an optimal division of the samples into testing and design sets depends on the type of classifier, the dimensionality, and on the asymptotic PMC. The theoretical results mentioned in Sections III and IV can help to find a solution, but no complete procedure has yet been devised. From (28) we have the MSE of the error counting estimate:

$$MSE\left(\hat{P}(n^* - N_1 - N_2)\right) = \frac{\hat{P}\left(1-\hat{P}\right)}{n^* - N_1 - N_2}. \tag{42}$$

For parametric classifiers, and assuming $N_1 = N_2 = N$, (16) yields

$$EP_N - P_\infty = t_\alpha(P_\infty, p)\frac{1}{N}, \tag{43}$$

where the coefficient $t_\alpha(P_\infty, p)$ depends on the classifier α, dimensionality, and P_∞ (Table II). For practical use of this methodology some prior guesses about the value of P_∞ should be available. Let $n^* = 300$, $N_1 = N_2 = \frac{N}{2}$, $C_1 = C_2 = 1$, the dimensionality $p = 8$, and assume the linear classifier is employed with $P_\infty = 0.1$. From a table in [45], we obtain $EP_N/P_\infty = 1.18$ when $N_1 = N_2 = 40$ and find $t_\alpha(0.1, 8) = (EP_N - P_\infty)(N_1 + N_2) = 0.018 \cdot 80 = 1.44$. Then the loss function (41) is

$$LOSS(N) = \frac{t_\alpha}{N} + \frac{\sqrt{P(1-P)}}{\sqrt{300-N}} = \frac{1.44}{N} + \frac{0.3}{\sqrt{300-N}}, \tag{44}$$

which attains a minimum near $N \approx 140$ and therefore $N_1 = N_2 = 70$, $n_t = 160$. The optimal balance for other values of P_∞ is found in a similar way.

VII. Discussion

We have reviewed a number of theoretical results available in statistical pattern recognition which highlight the difficulties caused by finite numbers of training and test samples. When we have a simple classification rule, such as the Fisher linear discriminant function and the number of features is not too large (5–10) then even a small number of training samples (50–100 observations per class) is sufficient to design a reliable decision rule. However, when the decision rule is complex, the number of features is large, or the number of training and test samples are small, then there are several design issues which need careful attention. We have presented theoretical results which show the existence of "curse of dimensionality" (peaking in classification accuracy as number of features are increased), and provided expressions which determine the bias and the variance of error rate estimators in finite sample situations.

Theoretical results regarding the classification accuracy and error rate estimates have been derived in the literature only for specific statistical models, usually multivariate Gaussian class-conditional densities. In practical problems, these models are often neither known nor appropriate. So, the theoretical results can only provide guidance to the pattern recognition system designer in selecting an appropriate methodology. The final design should be based on empirical results obtained by comparing competing algorithms.

Based on our experience and the available results, we provide the following set of recommendations to designers of pattern recognition systems.

1) Finite number of training samples require the designer to pay careful attention in selecting several design parameters, including a) number of features used in decision making, b) number of neighbors in a k-NN decision rule, and c) width of the Parzen window in density estimation.
2) With the availability of powerful desktop workstations and modern statistical packages, it is fairly easy to evaluate competing classifiers, investigate different feature selection and extraction methods, and estimate a classifier's performance using compute-intensive methods such as leave-one-out and bootstrap. There is no need to discard certain algorithms just because they demand too much computation; the choice of feature selection, classification, and error estimation procedures in the design of a pattern recognition system is done "off-line."
3) The estimate of the classification error depends on the particular training and test samples used, so it is a random variable. One should, therefore, investigate the bias and the variance of the error rate estimates. In particular, one should ask whether enough test samples were used to evaluate the classifier, and were the test samples different from the training samples?
4) Special attention should be paid to the problem of feature subset selection. In practice, we try to find that subset which gives the smallest classification error. But, as mentioned earlier, the estimated classification error can have a large variance which causes optimistic bias and will make it difficult to select the "best" subset.
5) The large error rate of a classifier can usually be attributed to the inherent difficulty of the classification problem. However, in finite sample situations, the following factors may also degrade the performance of a classifier: a) small number of training samples, b) large number of features, c) complexity of the classification rule (quadratic discriminant function versus linear discriminant function),

d) presence of outliers if a parametric classifier is being used, and e) inappropriate window width for a classifier involving nonparametric kernel density estimates.

Several important problems related to the design of a pattern recognition system when small number of samples are available remain open. For example, the influence of training sample size on the classification performance of nonparametric and piecewise linear classifiers is generally not known. With a resurging interest in artificial neural networks, these classifiers have become popular. There is also a need to establish a standard database, so that empirical studies can be carried out meaningfully. By default, most empirical studies are carried out on Gaussian data.

In conclusion, small sample effects make the problem of designing a pattern classification system very difficult, and these effects should not be ignored in practice.

References

[1] R. A. Abusev and Y. P. Lumelskij, "Unbiased estimators and classification problems for multivariate normal populations," *Theor. Prob. and Appl.*, vol. 25, pp. 381–389, 1980 (in Russian).

[2] S. A. Aivazian, V. M. Buchstaber, I. S. Yenyukov, and L. D. Meshalkin, "Applied statistics: Classification and reduction of dimensionality," *Finansy i Statistika* (Reference Edition), Moscow, 1989 (in Russian).

[3] B. G. Batchelor and D. J. Hand, "Pattern recognition competition," in *Proc. 3rd Int. Conf. Pattern Recognition*, Coronado, 1976, pp. 315–321.

[4] M. Ben-Bassat, "Use of distance measures, information measures and error bounds in feature evaluation," in *Handbook of Statistics*, vol. 2, P. R. Krishnaiah and L. N. Kanal, Eds. Amsterdam, The Netherlands: North-Holland, 1982, pp. 773–791.

[5] L. Breiman, J. Friedman, R. A. Olsen, and C. J. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.

[6] Y. D. Broffitt, "Nonparametric classification," in *Handbook of Statistics*, vol. 2, P. R. Krishnaiah and L. N. Kanal, Eds. Amsterdam, The Netherlands: North-Holland, 1982, pp. 139–168.

[7] B. Chandrasekaran and A. K. Jain, "On balancing decision functions," *J. Cybern. Inform. Sci.*, vol. 2, pp. 12–15, 1979.

[8] L. Devroye and T. J. Wagner, "Nearest neighbor methods in discrimination," in *Handbook of Statistics*, vol. 2, P. R. Krishnaiah and L. N. Kanal, Eds. Amsterdam, The Netherlands: North-Holland, 1982, pp. 193–198.

[9] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[10] B. Efron, "The efficiency of logistic regression compared to normal discriminant analysis," *J. Amer. Statist. Assoc.*, vol. 70, pp. 892–898, 1975.

[11] I. S. Enukov, "A choice of a set of measurements with maximal discriminating power in the case of limited learning sample size," in *Multivariate Statistical Analysis in Social-Economic Research*. Moscow, USSR: Nauka, 1974, pp. 394–397 (in Russian).

[12] D. M. Foley, "Considerations of sample and feature size," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 618–626, 1972.

[13] K. Fukunaga, "Statistical pattern recognition," in *Handbook of Pattern Recognition and Image Processing*, T. Y. Young and K. S. Fu, Eds. New York: Academic, 1986, pp. 3–32.

[14] K. Fukunaga and L. D. Hostetler, "Optimization of K-nearest-neighbor density estimates," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 320–326, 1973.

[15] S. Geiser, "Posterior odds for multivariate normal classifications," *J. Roy. Statist. Soc. B*, vol. 21, no. 1, pp. 69–76, 1964.

[16] N. Glick, "Additive estimators for probabilities of correct classification," *Pattern Recog.*, vol. 10, no. 3, pp. 211–222, 1978.

[17] M. Goldstein and W. R. Dillon, *Discrete Discriminant Analysis*. New York: Wiley, 1978.

[18] V. Grabauskas, Inst. Math. Cybern., Acad. Sci., Lithuania, personal communication, 1983.

[19] D. Griškevičius and Š. Raudys, "On the expected probability of the classification error of the classifier for discrete variables," in *Statistical Problems of Control*, issue 38, Š. Raudys, Ed. Vilnius, USSR: Inst. Math. Cybern. Press, 1979, pp. 95–112 (in Russian).

[20] D. J. Hand, "Recent advances in error rate estimation," *Pattern Recog. Lett.*, vol. 5, pp. 335–346, 1986.

[21] A. K. Jain, R. C. Dubes, and C. C. Chen, "Bootstrap techniques for error estimation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-9, no. 9, pp. 628–636, 1987.

[22] A. K. Jain and B. Chandrasekaran, "Dimensionality and sample size considerations in pattern recognition practice," in *Handbook of Statistics*, vol. 2, P. R. Krishnaiah and L. N. Kanal, Eds. Amsterdam, The Netherlands, North-Holland, 1982, pp. 835–855.

[23] A. K. Jain and M. D. Ramaswami, "Classifier design with Parzen windows," in *Pattern Recognition and Artificial Intelligence*, E. S. Gelsema and L. N. Kanal, Eds. Amsterdam, The Netherlands: Elsevier, 1988, pp. 211–228.

[24] A. K. Jain and W. G. Waller, "On the optimal number of features in the classification of multivariate Gaussian data," *Pattern Recog.*, vol. 10, pp. 365–374, 1978.

[25] L. Kanal, "Patterns in pattern recognition 1968–1974," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 697–722, 1974.

[26] L. Kanal and B. Chandrasekaran, "On dimensionality and sample size in statistical pattern classification," *Pattern Recog.*, vol. 3, pp. 238–255, 1971.

[27] D. G. Keehn, "A note on learning for Gaussian properties," *IEEE Trans. Inform. Theory*, vol. IT-11, no. 1, pp. 126–131, 1965.

[28] J. Kittler, "Feature selection and extraction," in *Handbook of Pattern Recognition and Image Processing*, T. Y. Young and K. S. Fu, Eds. New York: Academic, 1986, pp. 60–83.

[29] P. A. Lachenbruch and R. M. Mickey, "Estimation of error rates in discriminant analysis," *Technometrics*, vol. 10, no. 1, pp. 1–11, 1968.

[30] P. A. Lachenbruch, C. Sneeringer, and L. T. Revo, "Robustness of the linear and quadratic discriminant functions to certain types of non-normality," *Commun. Statist.*, vol. 1, no. 1, pp. 39–56, 1972.

[31] G. S. Lbov, "Logical functions in the problems of empirical prediction," in *Handbook of Statistics*, vol. 2, P. R. Krishnaiah and L. N. Kanal, Eds. Amsterdam, The Netherlands: North-Holland, 1982, pp. 479–491.

[32] T. Lissack and K. S. Fu, "Error estimation in pattern recognition via L-distance between posterior density functions," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 34–45, 1976.

[33] G. J. McLachlan, "The bias of the apparent error rate in discriminant analysis," *Biometrika*, vol. 63, pp. 239–244, 1976.

[34] ——, "Assessing the performance of an allocation rule," *Comput. Math. Applicat.*, vol. 12A, pp. 261–272, 1976.

[35] ——, "The efficiency of Efron's 'bootstrap' approach to error estimation in discriminant analysis," *J. Stat. Comput. Simulation*, vol. 11, pp. 273–279, 1980.

[36] ——, "Error rate estimation in discriminant analysis: Recent advances," in *Advances in Multivariate Statistical Analysis*, A. K. Gupta, Ed. Dordrect, The Netherlands: Reidel, 1987, pp. 233–252.

[37] L. Miroshnichenko, "Comparison of algorithms for selecting the best feature set in pattern recognition," in *Statistical Problems of Control*, issue 93. Vilnius, USSR: Inst. Math. Cybern. Press, 1990, pp. 78–91 (in Russian).

[38] T. Y. O'Neill, "The general distribution of the error rate of a classification procedure with application to logistic regression discrimination," *J. Amer. Statist. Assoc.*, vol. 75, pp. 154–160, 1980.

[39] K. W. Pettis, T. A. Bailey, A. K. Jain, and R. C. Dubes, "An intrinsic dimensionality estimator from near-neighbor information," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-1, no. 1, pp. 25–37, 1979.

[40] V. Pikelis, "Analysis of learning speed of three linear classifiers," Ph.D. dissertation, Inst. Phys. Math., Vilnius, pp. 1–136, 1974 (in Russian).

[41] Š. Raudys, "On the problems of sample size in pattern recognition," in *Proc. 2nd All-Union Conf. Statistical Methods in Control Theory*, Moscow, USSR: Nauka, 1970, pp. 64–67 (in Russian).

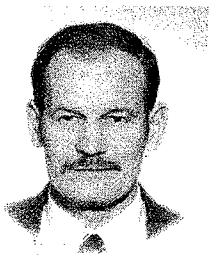
[42] Š. Raudys, V. Pikelis, and K. Juškevičius, "Experimental comparison of thirteen classification algorithms," in *Statistical Problems of Control*, issue 11, Vilnius, USSR: Inst. Phys. Math. Press, 1975, pp. 35–80 (in Russian).

[43] Š. Raudys, "Comparison of the estimates of the probability of misclassification," in *Proc. 4th Int. Conf. Pattern Recognition*, Kyoto, Japan, Nov. 1978, pp. 280–282.

[44] ——, "Determination of optimal dimensionality in statistical pattern classification," *Pattern Recog.*, vol. 11, pp. 263–270, 1979.

[45] Š. Raudys and V. Pikelis, "On dimensionality, sample size, classification error, and complexity of classification algorithm in pattern

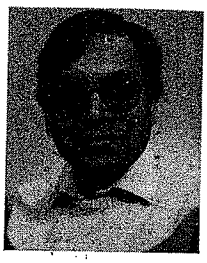
recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-2, no. 3, pp. 242–252, 1980.
[46] Š. Raudys, "The influence of sample size on classification performance," in *Statistical Problems of Control,* issue 66. Vilnius, USSR, Inst. Math. Cybern. Press, 1984, pp. 9–42 (in Russian).
[47] Š. Raudys and V. Vaitukaitis, "Methods to estimate the probability of misclassification," in *Statistical Problems of Control,* issue 66. Vilnius, USSR: Inst. Math. Cybern. Press, 1984, pp. 43–65 (in Russian).
[48] Š. Raudys, "On the accuracy of a bootstrap estimate of the classification error," in *Proc. 9th Int. Conf. Pattern Recognition,* Rome, Italy, Nov. 1988, pp. 1230–1232.
[49] Š. Raudys, V. Pikelis, and D. Stasaitis, "The effects of the number of initial and final features, the dependence between the features and the type of a classification rule on the accuracy of feature selection," *Pattern Recog. Artificial Intell.*, 1990, submitted for publication.
[50] J. W. Sayre, "The distribution of actual error rates in linear discriminant analysis," *J. Amer. Statist. Assoc.*, vol. 75, pp. 201–205, 1980.
[51] I. K. Sethi and G. P. R. Sarvarayudu, "Hierarchical classifier design using mutual information," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-4, pp. 441–445, 1982.
[52] M. Siotani, "Large sample approximations and asymptotic expansions of classification statistics," in *Handbook of Statistics,* vol. 2, P. R. Krishnaiah and L. N. Kanal, Eds. Amsterdam, The Netherlands: North-Holland, 1982, pp. 61–100.
[53] M. Skurikhina, "Effect of the kernel form on the quality of nonparametric Parzen window classifier," in *Statistical Problems of Control,* issue 93. Vilnius, USSR: Inst. Math. Cybern. Press, 1990 (in Russian).
[54] G. T. Toussaint, "Bibliography on estimation of misclassification," *IEEE Trans. Inform. Theory,* vol. 20, pp. 472–479, 1974.
[55] N. Vanichsetakul, "Tree structured classification via recursive discriminant analysis," Ph.D. dissertation, Univ. Wisconsin, 1986.
[56] V. N. Vapnik, *Recovery of Dependencies from Empirical Data.* New York: Springer-Verlag, 1982.
[57] C. T. Wolverton and T. J. Wagner, "Asymptotically optimal discriminant functions for pattern classification," *IEEE Trans. Inform. Theory,* vol. IT-15, no. 2, pp. 258–265, 1969.
[58] D. Zvirėnaitė, "Criteria for selecting the informative features in pattern recognition," in *Statistical Problems of Control,* issue 74. Vilnius, USSR: Inst. Math. Cybern. Press, 1986, pp. 76–103 (in Russian).



Sarunas J. Raudys was born in Kaunas, Lithuania, on February 24, 1941. He received the M.S. degree in electrical and computer engineering from Kaunas Polytechnical Institute in 1963, and the Candidate of Sciences and Doctor of Sciences degrees from the Institute of Mathematics and Cybernetics, Academy of Sciences, Lithuania, in 1969 and 1978, respectively.

He is currently Head of the Department of Data Analysis in the Institute of Mathematics and Cybernetics and Professor in the Department of Control Systems at Kaunas Polytechnical Institute, Lithuania. His current research interests include statistical pattern recognition, artificial neural nets, expert systems, machine learning, and data analysis methods.

Dr. Raudys is a member of the Classification Societies of the U.S.S.R. and France. He is an Associate Editor of the *Pattern Recognition Journal.* He has been a member of the Program Committee of INTERFACE-90 and other Soviet and international conferences.



Anil K. Jain (S'70–M'72–SM'86–F'91) received the B.Tech. degree from the Indian Institute of Technology, Kanpur, and the M.S. and Ph.D. degrees in electrical engineering from the Ohio State University.

He is a Professor in the Department of Computer Science at Michigan State University, East Lansing. His research interests are pattern recognition and computer vision. He has also served as Program Director of the Intelligent Systems Program at the National Science Foundation. He has been a consultant to a number of industrial organizations.

Dr. Jain is Editor-in-Chief of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. He is an Associate Editor of *Pattern Recognition Journal* and *Journal of Intelligent Systems* and is an Advisory Editor of *Pattern Recognition Letters.* He is the coauthor of *Algorithms for Clustering Data* (Prentice-Hall, 1988), has edited the book *Real-Time Object Measurement and Classification* (Springer-Verlag, 1988), and has co-edited the book *Analysis and Interpretation of Range Images* (Springer-Verlag, 1989).