Dimensionality and Sample Size Considerations in Pattern Recognition Practice

A. K. Jain* and B. Chandrasekaran**

1. Introduction

The designer of a statistical pattern classification system is often faced with the following situation: finite sets of samples, or paradigms, from the various classes are available along with a set of measurements, or features, to be computed from the patterns. The designer usually proceeds by estimating the class-conditional densities of the measurement vector on the basis of the available samples and uses these estimates to arrive at a classification function. Naive intuition suggests that if the dimensionality of the measurement vector is increased, then the classification error rate should generally decrease. In the case where the added measurements do not contribute in any way to classification, then the error rate should at least stay the same. For, after all, is not more information being utilized in the design? However, in practice quite often the performance of the classifier based on estimated densities improved up to a point, then started deteriorating as further measurements were added, thus indicating the existence of an optimal measurement complexity when the number of training samples is finite. The past decade has seen much research devoted to elucidating this phenomenon under various conditions [30].

The purpose of this paper is to discuss the role which the relationship between the number of measurements (dimensionality of the pattern vector, or simply dimensionality) and the number of training patterns (sample size) plays at various stages in the design of a pattern recognition system. The designer of a pattern recognition system can (and should) pose the following basic question: For a given knowledge about the form of the underlying class-conditional densities and the availability of certain numbers of training samples, how many measurements should be used in designing the classifier? While no specific design equations are available, our review below shows that general guidelines can be used to clarify the situation, and help the designer be aware of several possible pitfalls.

*Research supported by NSF Grants ENG 76-11956, Ao1 and ECS 8007606.
**Research supported by AFOSR Grant 72-2325.
2. Classification performance

The most well-known example of the "curse of finite sample size" is the peaking in the classification performance as the number of measurements is increased for a fixed number of training samples. Consider a two-class pattern recognition problem, where a total of \( N \) measurements is made on each pattern. Let the prior probabilities of the two classes be equal for simplicity and let \( f_i(x) \) be the class-conditional density function of the measurement vector \( x \) from class \( C_i, i=1,2 \). Let us first consider the simple situation where the number of training samples equals infinity, that is, the class-conditional density functions are completely known. Then, the probability of correct recognition based on the optimal Bayes decision rule is given by

\[
P_d(N) = \left\{ \frac{P_i}{P_j} \right\} \left[ P_i \log \frac{f_i(x)}{f_j(x)} > 0 \text{ for } x \in C_i \right] + P_i \log \frac{f_i(x)}{f_j(x)} > 0 \text{ for } x \in C_j \right].
\]

It is well known that \( P_d(N) < P_d(N+1) \), or in the presence of perfect information, the classification accuracy would never decrease as the number of measurements is increased. Whether or not \( \lim_{N \to \infty} P_d(N) \) approaches unity, i.e., perfect discrimination is attained in the limit, is, however, a different issue and has been studied in [7, 17, 27], and more recently in [65]. For example, in [65] it is shown that given two classes with known densities \( f_i(x) \) and \( f_j(x) \), if

\[
\lim_{N \to \infty} \frac{\mathbb{E}_i \left[ \log f_i(x)/f_j(x) \right]^2}{\mathbb{E}_j \left[ \log f_i(x)/f_j(x) \right]^2} = \infty,
\]

then the probability of correct classification for objects from \( C_i \) tends to unity, where \( \mathbb{E}_i \) and \( \mathbb{E}_j \) denote, respectively, the expectation and variance with respect to the \( f_i \) distribution. A similar result holds for \( C_j \).

A more realistic and at the same time mathematically tractable pattern recognition problem involves the case where the form of the densities \( f_i(x) \) is known but parameter values are unknown. Let \( m \) be the number of samples available from class \( C_i, i=1,2 \), to train the recognizer. The class-conditional densities need to be estimated from these samples. In the Bayesian formulation, some a priori densities on the parameters of \( f_i(x) \) are assumed, and through the sample set \( y_i, f_i(y_i) \), can be calculated. On the other hand, one can arrive at maximum likelihood estimates of the density function (by, say, first obtaining the maximum likelihood estimates of the unknown parameters and then substituting these estimated values for the true parameters in \( f_i(x) \)) if one does not want to concern oneself with a priori densities on parameters. These estimated density functions are then used in the Bayes’ decision function.

Given the knowledge of a priori densities on the unknown parameters, the Bayesian method in the design of a classification system is optimal, that is, there does not exist any other decision rule which has a higher recognition accuracy. The fact that a priori densities are required to be known and the complexity involved in computing the a posteriori densities even in the commonly occurring case of multivariate Gaussian densities with unknown mean vectors and covariance matrices [40] restrict and limit the scope of the Bayesian method. Therefore, in most practical applications, a suboptimal procedure, such as using maximum likelihood estimates of the parameters in place of their true values, is preferred. In the statistical literature the Bayesian method is often referred to as the predictive method, and the terms estimative procedure and plug-in rule are used to denote the method in which the unknown parameter is replaced by its estimate. Recent investigation by Atchison [2] and Atchison et al. [3] concerns the conditions under which the predictive method of statistical discrimination has superior properties. Clearly, the classification performance depends on the estimation procedure used, and we need to study problems pertaining to the relationship between dimensionality and sample size in the context of the method of estimation.

Allais [4] pointed out an interesting relation between dimensionality, sample size and recognition accuracy in the linear prediction problem. He considered an unobservable random variable, \( y \), an observable measurement vector, \( x \), and a linear predictor of \( y \), \( g(x) \), represented as

\[
g(x) = c + w^T x
\]

where \( c \) is a scalar and \( w^T \) is a row vector. The performance of \( g(x) \) was evaluated in terms of its mean square error defined as

\[
e = \mathbb{E}[(y - g(x))^2]
\]

where \( \mathbb{E} \) denotes the expectation operator. Allais considered the case where the joint distribution of predictor \( y \) and measurement vector \( x \) was multivariate Gaussian. Since the parameters of this distribution were assumed not known, Allais considered a maximum likelihood predictor \( g(x) \) and derived its unconditional mean square error as

\[
E(e) = \begin{cases} 
\frac{m+1}{m} & \text{for } N < m - 2, \\
\frac{N - m - 2}{N} & \text{for } N = m - 1, \\
0 & \text{for } N > m,
\end{cases}
\]

where \( N \) is the number of observable measurements, \( m \) is the number of samples and \( \theta \) is the ideal mean square error assuming all parameters of the distribution are known. It can be easily verified from the above expression that for a fixed sample size the mean square error of the maximum likelihood predictor decreases at first, attains a minimum and starts increasing as the number of measurements is gradually increased. Results reported by Allais raised the following question: Is
the existence of a finite measurement complexity due only to finite sample size, or whether the kind of estimates (maximum likelihood vs. Bayesian) used had any relationship to it?

2.1. Optimal Bayes classification rule

Hughes [30] considered the behavior of a finite-sample Bayesian classifier with respect to increasing dimensionality. The model considered in [30] and [1] can be summarized as follows.

Let the measurement \( x \) take one of \( n \) possible values where \( n \) is the measurement complexity. Let \( a_i = P(x = x_i | c_1) \) and \( b_i = P(x = x_i | c_2) \) such that

\[
\sum_{i=1}^{n} c_i = \sum_{i=1}^{n} b_i = 1.
\]

If the pattern vector is an \( n \)-dimensional binary vector for instance, \( a \) would be \( 2^n \). The sets \( \{a_i\} \) and \( \{b_i\} \) specify a particular pattern recognition problem. If some a priori densities on \( \{a_i\} \) and \( \{b_i\} \) are assumed, and if one uses a Bayesian solution, then from available samples (say \( m, m_i \) in number from \( c_i \)) Bayesian estimates of \( \{a_i\} \) and \( \{b_i\} \) can be obtained, and the optimal decision function for \( x = x_i \) is then: decide class \( c_i \) if \( \hat{a}_i > \hat{b}_i \), class \( c_2 \) otherwise. Assuming a a priori distribution on \( \{a_i\} \) and \( \{b_i\} \) to be uniform in the respective simplices, Hughes [30, 1] computed a quantity \( F_s(n, m, m_i) \) that is the average probability of correct recognition over all problems generated by the assumed a priori densities on \( \{a_i\} \) and \( \{b_i\} \) for fixed sample sizes \( m, m_i \). This quantity increases with \( n \) up to a point \( n_{opt} \) called the optimal measurement complexity, then starts decreasing until as \( n \rightarrow \infty \), \( F_s \rightarrow 1 \). As the number of training samples increases, \( n_{opt} \) increases and in the limit as \( m_i \rightarrow \infty \) and \( m \rightarrow \infty \), \( n_{opt} \rightarrow \infty \), indicating the absence of peaking in the infinite sample case. Kain [36] extended the results of Hughes to the multi-class case.

The results of [30] and [1] showed that even if a Bayesian procedure was used which is optimal in the sense that it minimizes the probability of misclassification, one could get into trouble by using too many measurements when the number of training samples was finite. This led to a number of investigations [11, 12, 69] where attempts were made to find useful exceptions to this "curse of finite sample size" in the Bayesian context. For example, it was shown that if the measurements are independent and binary [11], or independent and quantized to \( k \) levels [12], or first-order nonstationary Markov dependent and binary [69], then there was no peaking of performance in the Bayesian context with respect to the measurement complexity for finite sample size. Kanal and Chandrasekaran [39] emphasized the relationship between the probability structure of the classification problem, sample size and dimensionality to explain the peaking in Hughes' model and the monotonicity of \( F_m \) in the model of Chandrasekaran [11]. They stated that for a given sample size the optimum measurement complexity as well as the maximum probability of correct recognition increases with the increased structure in the classification problem.

The peaking of classification accuracy when the Bayesian procedure is used as reported in [30] needs some explanation due to the optimality of the Bayesian procedure. This apparent paradox was resolved recently by Van Campenhout [64] and Waller and Jain [70], but was already implicit in the earlier work of Lindley [44, 45]. They show that even with a finite set of training samples the performance of a Bayesian classifier can not be degraded by increasing the number of features as long as the old set of features is recoverable from the new set of features [70] or the two classification problems (as a result of increasing the number of features) are comparable [64]. The peaking phenomenon reported by Hughes [30] can be attributed to incompressible prior densities. To be more specific, let us consider two classification problems in the context of Hughes' model consisting of \( n \) and \( n' \) measurement states, and let the corresponding prior densities on the unknown parameters be \( g(\theta) \) and \( g(\theta') \), respectively, for class \( c_j, i = 1,2 \). Hughes assumed these prior densities to be uniform over the \( n \)- and \( n' \)-dimensional simplices respectively, and for a fixed sample size \( m_i \) from class \( c_j, i = 1,2 \), he computed the mean recognition accuracies \( P_e(n, m, m_i) \) and \( P'_e(n', m, m_i) \). These mean recognition accuracies involve two types of averaging—first an average is taken over all of the possible sets of random samples of size \( m_i \) from class \( c_j \) and then the resulting error rate is averaged over all possible problems generated by the prior densities. Both Van Campenhout [64] and Waller and Jain [70] point out that \( F_s(n, m, m_i) \) and \( F_s(n', m, m_i) \) corresponding to the two classification problems, are incompressible due to incompressible priors. That is, if the prior density \( g(\theta) \) for parameters in an \( n \)-measurement-states problem is uniform, then \( g(\theta') \) for parameters in a problem with \( n' \) number of states is not uniform as assumed in [30, 12].

In summary, the results of [64, 70] have proved that if the designer of a pattern recognition system would be willing to take the Bayesian approach, then even for a finite number of training samples the number of measurements can be increased arbitrarily without the fear of peaking in the average recognition accuracy. In other words, taking a Bayesian approach results in the optimum measurement complexity being equal to infinity irrespective of the number of training samples available. This does not imply that using Bayesian approach would always lead to \( F_s(n, m, m_i) \rightarrow 1 \), or that by the addition of arbitrarily large number of measurements in the classifier design one would get perfect discrimination. Particular cases in which perfect discrimination is possible in the limit as the number of measurements is increased are investigated in [11, 12, 69]. Even though one is guaranteed not to do worse by taking more measurements in the Bayesian context, the cost and complexity due to additional measurements may not be worth the slight improvement (if there is any) in the recognition accuracy. Thus it will be interesting to know the rate of convergence of \( F_m \) and the conditions for perfect discrimination as the number of measurements is increased. What role does the structure of the underlying classification problem play in the perfor-
A. K. Jain and R. Chandrasekaran

mance of a Bayesian classifier? Waller and Jain [69] demonstrate that for a given sample size and measurement complexity, $P_e$ increases as the problem becomes more structured.

So far, in the Bayesian context, we have talked only about the average performance over all possible problems generated by the a priori densities. However, what about the performance $P_e$ in an individual problem (a specific set of parameters), even though the parameter estimates are Bayesian for the given a priori densities? This is the kind of performance in which a designer of a recognition system is really interested. In this case the notion of ‘optimality’ of the estimate does not enter, since the optimality of the Bayesian estimates is assured by averaging over a problem space generated by the a priori densities. In [13], [14] and [65], examples are given that show that peaking is indeed possible in this situation. More generally, not much is known about the conditions for perfect discrimination or the convergence of $P_e$ as the number of measurements is increased. If the measurements are independent, then the sufficient conditions of [13], [14] and [65], can be used to determine if, for a specific problem, perfect discrimination is possible in the limit. But in order to check for these conditions one must know the true parameters of the class-conditional densities, which, if available, would obviate the need for estimation in the first place. However, these conditions do illustrate the fact that a finite sample size imposes greater constraints on the measurement parameters to be ‘good’. For example, if the measurements are independent and binary with parameters $p_i = P(x_i = 1|c_i)$ and $q_i = P(x_i = 0|c_i)$, $i = 1, \ldots, N$, then, if one has infinite number of samples, a good measurement is one for which $|p_i - q_i| > \delta > 0$, a condition which is not sufficient for perfect discrimination for the finite sample case.

2.2. Suboptimal classification rules

The average recognition accuracy of a Bayesian classifier, as shown in [65, 70] will never decrease as the number of measurements is increased. However, when the approach to classifier design is non-Bayesian—typically in such a situation the parameters of the distributions may be estimated by say maximum likelihood methods—then peaking becomes theoretically possible. There is no longer any obvious natural notion of ‘optimality’, i.e., while the original Bayes’ rule is optimal, the decision rule that results by substituting the maximum likelihood estimates of the parameters is no longer optimal, and this is often the cause of peaking observed in the performance of many practical pattern classifiers. In some sense the errors caused by the nonoptimal use of added information outweigh the advantages of extra information. The mechanism behind this self-defeating behavior is the subject of this section.

Classification problems involving multivariate Gaussian densities have received the most attention in the literature both on statistics and on pattern recognition. This is understandable due to the ease in mathematical analysis and because class-conditional densities of many real-world classification problems can be reasonably approximated as multivariate Gaussian. Let us consider two equiprobable pattern classes which are represented by multivariate Gaussian densities with a common $N \times N$ non-singular covariance matrix $\Sigma$ and mean vectors $\mu_1$ and $\mu_2$. Assuming that these parameters ($\mu_1$, $\mu_2$ and $\Sigma$) are known, the decision rule which results in minimum classification error is a linear discriminant function given by: decide that vector $x$ belongs to class $c_1$ if

$$x^T \Sigma^{-1}(\mu_1 - \mu_2) > (1/2)(\mu_1 + \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2),$$

otherwise decide class $c_2$. It is well known that the probability of error using this discriminant function is related to the Mahalanobis distance $D_{\Sigma}^2$ between the two populations, given as

$$D_{\Sigma}^2 = (\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2).$$

The larger the Mahalanobis distance, the smaller the probability of error. Usually the parameters $\Sigma$, $\mu_1$, and $\mu_2$ are not known and the following estimates of these parameters based on $m_1$ training samples from class $c_1$ are commonly used:

$$\hat{\mu}_1 = \frac{1}{m_1} \sum_{j=1}^{m_1} x_j^{(i)}, \quad i = 1, 2,$$

$$\hat{\Sigma} = \frac{1}{(m_1 + m_2 - 2)} \sum_{i=1}^{m_1} (x_j^{(i)} - \hat{\mu}_1)(x_j^{(i)} - \hat{\mu}_1)^T,$$

where $\hat{\mu}_1$ is the sample mean of class $c_1$ and $\hat{\Sigma}$ is the pooled unbiased estimate of the common covariance matrix $\Sigma$. In the above expressions $x_j^{(i)}$ refers to the $j$th training sample from class $c_i$. In this situation, the most commonly used decision rule is based on the $W$ statistic proposed by Wald [5, 68] where

$$W = x^T \Sigma^{-1}(\hat{\mu}_1 - \mu_1) - (1/2)(\hat{\mu}_1 + \mu_1)^T \Sigma^{-1}(\hat{\mu}_1 - \mu_1).$$

The decision rule is to assign a feature vector $x$ to class $c_1$ if $W > 0$, otherwise assign it to class $c_2$. Clearly, this decision rule is suboptimal. Various authors have given the exact distribution of $W$ statistic and expressions for the associated probability of misclassification [9, 10, 28, 34, 35, 50, 57, 58, 59, 60]. Unfortunately, if the two classes have unequal covariance matrices (the resulting discriminant function is quadratic), then the determination of expected probability of misclassification is extremely difficult and to our knowledge no such expression exists.

Now that we have defined a commonly occurring classification problem involving two multivariate Gaussian distributions with unknown mean vectors and (common) covariance matrix, we are ready to discuss the role of dimensionality and sample size in the performance of a suboptimal classifier. In fact, almost all of the work which has been done pertaining to the relationship between dimensionality, sample size and probability of misclassification is in the context of the above mentioned model.
Rao [52] first illustrated the dangers of using too many measurements in a classification problem involving two populations some thirty years ago. He stated that “It does not seem to be, always, more the better...”. It is unfortunate that pattern recognition community is not aware of this pioneering work by Rao. The example used by Rao to illustrate this problem involved discrimination between Indian and Anglo-Indian skeletons based on two measurements—lengths of Femur and Humerus. Rao took 20 samples of Indian skeletons and 27 samples of Anglo-Indian skeletons and used the estimate of the Mahalanobis distance \( D^2 = (\bar{p}_1 - \bar{p}_2)S^{-1}(\bar{p}_1 - \bar{p}_2) \) to test if the separation between the two populations is significant (at 5% level). It was found that while the two populations were significantly different when only a single measurement (either Femur length or Humerus length) was used, there was no significant separation when both the measurements were used. Rao provides a test to determine whether the addition of \( q \) more measurements to an existing set of \( N \) measurements increases the distance between the two populations and concludes that if the Mahalanobis distance \( D^2 \) increases proportionately with the number of measurements, then, except in situations where the total number of samples is very small, the addition of extra measurements does not result in a loss in discrimination.

More recently Jain and Waller [33] have studied the peaking phenomenon in an effort to relate the optimum number of measurements to the number of available training samples and the Mahalanobis distance between the two populations. Their results can be related to those obtained by Rao [52]. For a classification problem involving two equiprobable multivariate Gaussian densities with a common covariance matrix they use the asymptotic expansion of the average probability of error (good up to order \( 1/m^3 \)) derived in [59] to conclude the following (of course, we are assuming that \( S \) and \( S \) given earlier are being used):

1. The minimal increase in the Mahalanobis distance needed to keep the same error rate when a measurement is added to a set of \( N \) features is

\[
\log \frac{\tilde{S}^2}{\tilde{S}^2/(2m - 3 - N)}
\]

where \( m \) is the number of training samples per class. In order to avoid peaking, \( \tilde{S}^2 > \tilde{S}^2/(2m - 3 - N) \). Note that increasing the sample size decreases the value of \( \tilde{S}^2 \) and in the limit as \( m \to \infty \), \( \tilde{S}^2 \to 0 \).

2. If the Mahalanobis distance is proportional to the number of measurements or, equivalently, if all the features are equally good, then the peaking in the performance of the classifier is not a real problem because \( N_m = m - 1 \).

In order to study the effect of the structure of the covariance matrix on the optimum number of measurements, Jain and Waller [33] considered three types of Toeplitz matrices for \( \Sigma \). However, in their analysis, the classifer did not incorporate knowledge of the form of the covariance matrix, and thus they used the general covariance matrix estimator \( S \) in the \( W \) statistic. As was expected, the performance of a classifier improved if it could utilize the knowledge about the structure of the covariance matrix, as was recently con-

firmed by Morgera and Cooper [48]. They introduce the notion of "effective sample size" to show that the constrained Toeplitz estimator provides better performance at a sample size \( m \) for which the generalized estimator has poor performance. In other words, for the same performance a classifier using constrained Toeplitz estimates requires fewer samples than if the generalized estimator is used, assuming that the true covariance matrix is of Toeplitz form. Interestingly enough this reduction in required sample size (to maintain the same performance) is an increasing function of the dimensionality \( N \).

Several recent investigations have demonstrated the relationship between dimensionality, sample size and recognition accuracy based on Monte Carlo simulations. Bourdon et al. [8] show that if the decision rule is in the form of a linear discriminant function, then a subset of measurements can yield better classification results than the full set of measurements. A real-world example they considered to demonstrate this consisted of data from a 12-channel sensor used in remote sensing of agricultural crops (soybeans and corn) obtained from NASA. If only thirty training samples are available from each of the two classes, then the expected probability of misclassification is lowest when only six measurements are used out of a possible total of twelve. The number of training samples per class must exceed one hundred before all twelve measurements can be used safely.

Van Ness and Simpson [67] and Van Ness [66] confirm the results of [33, 52] namely the Mahalanobis distance between the two populations must increase as the number of measurements is increased for the classification performance to stay at least the same. However, they do not provide any explicit expression for this increase as a function of sample size, dimensionality and Mahalanobis distance. Van Ness and Simpson [67] also compare experimentally the classificatory power of five different discriminant functions (in the order of assuming less and less knowledge about the true underlying distributions which were multivariate Gaussian: linear with unknown mean vectors and known covariance matrices, linear with unknown mean vectors and unknown but common covariance matrices, quadratic with unknown mean vectors and unknown covariance matrices and finally two non-parametric decision rules involving Parzen window estimates of the density function based on Gaussian and Cauchy window functions). A surprising result of this comparison was that the two non-parametric decision rules performed better than the linear and quadratic discriminant functions (with unknown covariance matrices) even when the dimensionality was small.

Results reported by Raudys [53] are similar to those in [8, 67]. Like Van Ness and Simpson [67], Raudys uses Monte Carlo simulations to generate tables showing the relationship between sample size, dimensionality, classification accuracy and complexity of the classification rule. As is understandable, this table depends on the true underlying class-conditionanl densities, which in the case of Raudys was assumed to be multivariate Gaussian with a common identity covariance matrix. An important guideline proposed by Raudys is that the number of training samples required to achieve a given recognition accuracy should increase linearly with the number of measurements for linear discriminant
functions, and should increase quadratically for quadratic discriminant functions. This is in general agreement with the acceptable code of good practice in pattern recognition design, namely the number of training samples must be five to ten times the number of measurements. Raudys suggests using his table to determine the optimal number of measurements for a given sample size, but this requires that the designer knows the true underlying distributions.

In an effort to isolate the effect of finite sample size on expected probability of error, Duan [21] derives the following expression for a two-class classification problem

\[
E(P_e) = P_e^* + E(e_1) + e_2
\]

where \(P_e^*\) is the optimal Bayes error and

\[
e_i = \frac{1}{2} \int \left| \tilde{f}_i(x) - f_i(x) \right| dx = 1 - \frac{1}{2} \int \min \{ \tilde{f}_i(x), f_i(x) \} dx,
\]

\(i = 1, 2\) and \(0 < e_i < 1\).

In the above expression, \(\tilde{f}_i(x)\) and \(f_i(x)\) are the estimates of the class-conditional densities which are used in the design of the classifier. Thus the quantities \(e_i\), \(i = 1, 2\), are the errors made in estimating the densities \(f_i(x)\), \(i = 1, 2\), respectively. It is clear that \(E(e_i)\) is a function of the sample size, dimensionality and the true underlying density. For Gaussian densities, Duan computes \(E(e_i)\) for different values of \(m\) and \(N\) using Monte Carlo runs, and shows that as dimensionality increases, more samples are needed to keep \(E(e_i)\) at some fixed value. This supports one of the main reasons given to explain peaking: for a fixed sample size, as the dimensionality increases, the extra discriminatory power provided by the added features is overcome, after a certain point, by the deterioration in the estimates of the densities. More work needs to be done to establish this point of crossover for underlying Gaussian densities with different parameter values.

So far, most of the results we have summarized deal with Gaussian densities in one way or another. Can some set of general conditions be obtained as a function of \(f_1(x), f_2(x), m_1, m_2\) and \(N\) such that for a two-class classification problem we are guaranteed to have perfect discrimination and monotonicity of expected probability of misclassification as the number of measurements is increased? Chandra [13, 14] provide a partial answer to the question raised above (see also [63] for some corrections to the results of [13]). To summarize the results of [14] and [63], let

\[f_1(x) = \prod_{i=1}^N f_1(x_i)\] \[f_2(x) = \prod_{i=1}^N f_2(x_i)\]

be the two class-conditional densities. The pattern vector \(x\) consists of \(N\) statistically independent measurements. Following the notation in [14] we use the following abbreviations:

\[d_i = \log \frac{f_i}{f_{i+1}}, \quad M_{i}(j) = E_{x_i} f_i \frac{d_i}{j}\]

and

\[P_i(j) = E_{x_{i+1}} \left( \frac{d_i - M_{i}(j)}{j} \right)^2\]

where \(j\) stands for estimated densities, \(c_i\) and \(c_{i+1}\) are the two classes, \(E_{x_i}\) stands for expectation with respect to class \(c_i\), and \(E_{x_{i+1}}\) for expectation with reference to training data set \(X_i\). Now the probability of correct classification, given that \(c_i\) is the true class, can be expressed as

\[P(\sum d_i > 0 | c_i) = P\left( \frac{\sum d_i - \sum M_{i}(j)}{(\sum P_i(j))^{1/2}} > \frac{-\sum M_{i}(j) - \sum P_i(j)}{(\sum P_i(j))^{1/2}} \right).\]

All the summations above and below are from \(1\) to \(N\).

Using Chebyshev's inequality it can be shown that the probability of error, given \(c_i\), is

\[P(\sum d_i < 0 | c_i) \leq \frac{\sum P_i(j)}{(\sum M_{i}(j))^{1/2}}\]

Thus a sufficient condition for the probability of correct classification for elements of \(c_i\) to approach one as \(N \to \infty\) is

\[
\lim_{N \to \infty} \left( \frac{-1}{\sum M_{i}(j)} \right)^{1/2} = \frac{1}{\sum P_i(j)} = \infty.
\]

These two sufficient conditions (for \(j = 1, 2\)) are useful in determining whether perfect discrimination is possible in the limit, but they do not assure us of the monotonicity of the probability of correct recognition. However, if the Central Limit Theorem can be applied so that the random variable \(\tilde{d}_i\) can be viewed to have a Gaussian distribution, then the two sufficient conditions also become necessary. Consequently, as the quantity \(\sum M_{i}(j) / (\sum P_i(j))^{1/2}\) increases by the addition of measurements, the probability of correct classification, given \(c_i\), increases correspondingly. Similar arguments hold for the probability of correct classification, given \(c_{i+1}\). While the verification of the conditions for the applicability of the Central Limit Theorem are not easy, let us consider the following example where we know them to be true.

Let the measurements be independent, continuous and normally distributed, \(f_1(x_i) = \mathcal{N}(\theta_i, 1)\) and \(f_2(x_i) = \mathcal{N}(\xi_i, 1)\) where \(\mathcal{N}(\mu, \sigma^2)\) is the normal density.
with mean $\mu$ and variance $\sigma^2$. If $\hat{\theta}$ and $\hat{\phi}$ are the maximum likelihood estimates based on $m$ training samples per class, then it can be shown that

$$M^0 = M^0(\hat{\theta}, \hat{\phi})^2,$$

$$\nu^0 = \nu^0 = K_0(m)(\hat{\theta} - \phi_i)_i^2 + K_0(m),$$

where $K_0$ and $K_2$ are positive for all positive $m$. Thus, the necessary and sufficient condition for perfect discrimination and monotonicity of recognition accuracy is that $\Delta_x^2 = E(\hat{\theta} - \phi_i)_i^2$ is of order $\nu^0/N$ as $N \to \infty$. Note that in this example the covariance matrix is the identity matrix which is known by the classifier whereas in the models of Rao [52] and Jain and Waller [33] the common covariance matrix has to be estimated. This explains why, in the less structured models of [52, 33], the Mahalanobis distance $\Delta_x^2$ is required to increase proportionally to $N$ to avoid peaking while in the above example it is sufficient that $\Delta_x^2$ be of order $\nu^0$. We can only hypothesize that if the two covariance matrices are unequal and have to be estimated, then $\Delta_x^2$ must be of order $N^{1/2}$ to avoid peaking.

The conditions derived by Chandrasekaran and Jain [14] are, as mentioned, for the case of statistically independent measurements. In [15] the same authors have generalized the conditions for the case of dependent measurements with arbitrary distributions. Let $d(N, x)$ be the Bayes' decision function such that $x \in c_i$ if $d(N, x) > 0$ and $x \in c_i$ otherwise, where the pattern vector $x$ has $N$ components, and let $\Delta d(N, x)$ be the classifier obtained by using estimates. Let further

$$M^0 = E_x \Delta d(N, x)$$

and

$$\nu^0 = E_x \Delta d(N, x)^2.$$

Then by arguments similar to those of Gaffey [27] and Van Ness [65], it can be shown that, if

$$\lim_{N \to \infty} M^0/\nu^0 = 1,$$

then $\lim_{N \to \infty} P_e = 1$ for elements of class $c_j$, $j \neq 1, 1, 2$. However, applying these conditions to actual cases will be more or less difficult depending upon the tractability of the underlying distributions. For the case of multivariate normal distributions with unknown mean vectors and known covariance matrices, more compact conditions for perfect discrimination are given in [15], and experimental and mathematical investigations of some aspects of performance as the dimensionality is increased are provided in [62] and [55].

2.3. Balancing decision functions due to unequal sample size

There is an interesting aspect of statistical pattern classification that has not received much attention: namely, the degradation in the classification performance when the numbers of training samples from various classes are substantially different. A few authors (Rao [52], Jain and Waller [33], and Levine, Littick and Salzberg [43]) have outlined the advantages of having equal numbers of samples. In the context of a linear discriminant function [52] showed that for given $\Delta_x^2, N$ and $m$, and using maximum likelihood estimates, it is more profitable to have $m_1 = m_2$. Note that $\Delta_x^2$ is often used as a feature selection criterion. Jain and Waller [33] used an asymptotic expansion to show that Okamoto's [50] probability of misclassification is minimum when $m_1 = m_2$ and that the degradation in the classification performance due to an unbalanced set of training samples is more severe for a large number of measurements. Levine et al. [43] show that for a nearest-neighbor decision rule, best results are obtained when $m_1 = m_2$. The above results suggest that if the designer can obtain only a small number of samples from one class, it is not necessary to compensate by taking large samples from the other class. In fact, Chandrasekaran and Jain [13] demonstrate a counter-intuitive phenomenon whereby discarding the excess samples from the class containing the greater number of samples is profitable.

What should a designer do when confronted with unequal numbers of training samples? In this situation the degrees of reliability associated with the estimates of the different class density functions are clearly different, and one has the intuitive feeling that this factor should somehow be taken into account, i.e., the decision function must be 'balanced' with respect to the different sample sizes. To be more concrete, while $f_1(x)$ and $f_2(x)$ might be individually the 'best' estimates of the density function, it is not at all clear that $d(x) = \log f_1(x) - \log f_2(x)$ is the best decision function to use if the sample sizes are substantially different. Perhaps a modification such as $\log W_1 f_1(x), m_1 - \log W_2 f_2(x), m_2$ where $W$ is a weighting function and $m_1$ and $m_2$ are the two sample sizes might perform better.

Kamimura and Watanabe [37] proposed the idea of a well-balanced adaptive decision function in the context of a perceptron convergence algorithm, when substantially different numbers of paradigms from the two classes are used. These authors proposed to adjust the position and orientation of the resulting hyperplane to reflect the different numbers of paradigms. In [37] the problem was posed in a non-probabilistic context, and the criterion was heuristic in nature. Chandrasekaran and Jain [15] observed that this question of balancing does not arise in the case where the classifier is Bayesian, i.e., prior distributions on unknown parameters are assumed and the classifier essentially determines $p(x | y)$ where $y$ denotes the sample set. In this approach the weighting is automatically incorporated in the posterior probabilities.

In [15] Chandrasekaran and Jain suggested that the criterion for weighting the estimated densities should be chosen so that the counter-intuitive phenomenon of doing better by discarding excess samples from the class containing a greater number of samples, reported in [13], does not arise. Consider a two-class problem involving multivariate Gaussian densities where the common covariance matrix is known, and the mean vectors $\theta$ and $\phi$ are unknown. They showed that if maximum likelihood estimates of $\theta$ and $\phi$ are used based on $m_1$ and $m_2$ samples, respectively, then a balanced decision function based on the above criterion is

$$\hat{d}_x(x) = m_1/m_1 - m_1/m_2 + \hat{d}(x)$$
where
\[ d(x) = -(x - \hat{\theta})^T \Sigma^{-1} (x - \hat{\theta}) + (x - \hat{\phi})^T \Sigma^{-1} (x - \hat{\phi}). \]

Note that \( \tilde{d}(x) \) is the usual linear decision function. In case \( m_2 = m_3 = \tilde{d}(x) = \hat{d}(x) \), and as \( m_1 \) and \( m_2 \) approach infinity, then \( \tilde{d}(x) \) approaches \( d(x) \), where \( d(x) \) is the Bayes' decision surface for this problem. An interesting aspect of this particular weighting is that \( K_1 \tilde{d}(x) = \tilde{d}(x) \), i.e., the weighting results in the expected value (over the samples) of the estimated decision surface being the same as the Bayes' decision surface.

The results of [13] provide an efficient way to utilize information provided by unequal number of samples when a non-Bayesian decision rule is employed. These results need to be generalized to a larger class of distributions and estimates.

2.4. Weaker conditions for perfect discrimination

Continuing with the case of statistically independent measurements, the sufficient conditions for perfect discrimination were (Section 2, the first part) that certain quantities (corresponding to some weighted distances between the class distributions) approached \( \infty \) as the number of measurements was increased to \( \infty \). Of course these conditions were derived for the case of a particular (classical) approach to the design of a classifier; i.e., the decision function is \( \Sigma \tilde{d} \) (see before for the notation). Schaafmay and Steeneman [5] (see also the article by Schaafmay in this handbook) consider an interesting question: Do there exist other classification rules for which perfect discrimination can be achieved under weaker conditions, e.g., that the quantity
\[
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{N} \frac{M_i^{(i)}}{\left( \sum_i M_i^{(i)} \right)^{1/2}}
\]
is simply \( > 0 \) rather than approaching \( \infty \)? We have already seen that by weighting the estimated densities of each class by a function which reflects the sample size, we can avoid some undesirable consequences of differential sample sizes from different classes. Schaafmay and Steeneman similarly look for other classifiers (than the classical ones) from the viewpoint of achieving perfect discrimination under weaker conditions on the distance between densities.

They consider the familiar case of two classes, each with \( N \) independent and normally distributed measurements, \( f_1(x) = \mathcal{N}(\theta, 1) \) and \( f_2(x) = \mathcal{N}(\phi, 1) \). As before, \( \Delta^2 = \mathcal{N}(\theta - \phi)^2 \), and \( m_1 \) and \( m_2 \) are the sample sizes from the two classes. Consider the use of the decision function
\[
d = \frac{m_1}{1 + m_1} \sum_{i=1}^{N} \frac{1}{N} \| (x_i - \hat{\theta})^2 \|_1 + \frac{m_2}{1 + m_2} \sum_{i=1}^{N} \frac{1}{N} \| (x_i - \hat{\phi})^2 \|_1.
\]

Note that this rule weights the contribution of different measurements differently in addition to taking into account different sample sizes. They show that if
\[
\lim_{N \to \infty} \frac{1}{N} \Delta^2 > 0,
\]
then the probability of misclassification approaches 0. The important point is not the particular procedure or decision function used, but the existence of a class of procedures which can produce both perfect discrimination in the limit under weaker conditions on \( \Delta^2 \), and elimination of certain peculiarities associated with different sample sizes. This indicates that it might be quite fruitful to study the general problem of nonclassical discrimination rules to provide some protection from the "curse of finite sample size".

3. K-nearest neighbor procedures

The information contained in the \( K \) nearest neighbors of a sample pattern vector \( x \) has been used extensively at least in two areas of pattern recognition: classification of \( x \) into one of several possible categories, and estimating the density at point \( x \). The popularity of \( K \)-nearest neighbor procedures is due to their simplicity and because their applicability is not conditioned on the knowledge of the form of the underlying densities. Thus the designer may only have information of the following kind: There are \( C \) pattern classes with \( m_i \)-dimensional training samples from class \( c_i, i = 1, \ldots, C \). The designer who decides to use the nearest-neighbor procedure almost always faces the question: what value of \( K \) should be used?

Let us first consider the \( K \)-nearest-neighbor decision rule. It is well known [19, 20] that if the number of training samples is infinite, then
(i) the error rate of the nearest-neighbor rule is bounded from above by twice the Bayes rate, and
(ii) the performance of the \( K \)-nearest-neighbor rule improves monotonically with \( K \), and in the limit as \( K \) goes to infinity, this rule becomes identical to the optimal Bayes rule.

Unfortunately, no such statements can be made about the performance of the \( K \)-nearest-neighbor decision rule when only a finite number of training samples is available. Fix and Hodges [23] attempted to approximate the finite sample error rate of the \( K \)-nearest-neighbor decision rule and compare its performance with the linear discriminant function when \( K \leq 2 \) and \( K < 3 \). In practice it is quite often observed [6] that for a fixed \( m \) (number of training samples per class) and \( N \) (dimensionality), as \( K \) increases, the performance of the \( K \)-nearest-neighbor rule improves, attains a maximum and then starts deteriorating. Thus, we have another example of peaking of performance.

It is often recommended [20] that \( K \) should be only a small fraction of the number of samples. The relationship between \( K, m \) and \( N \) needs to be further investigated especially since the \( K \)-nearest-neighbor decision rule is so popular.
One common method to obtain nonparametric density estimates is to use the $K$-nearest-neighbor approach. Given $m$ $N$-dimensional samples, the density estimate $p_n(x)$ at point $x$ can be obtained as

$$p_n(x) = \left(K_n/m\right)/V_n$$

where $V_n$ is the volume of a region centered at $x$ which captures exactly $K_n$ samples. (The subscript $m$ is used to denote the dependence of various quantities on the sample size $m$.) The necessary and sufficient conditions for $p_n(x)$ to converge to $p(x)$ are that $\lim_{n \to \infty} K_n = \infty$ and $\lim_{n \to \infty} K_n/m = 0$ [46]. Empirical results [46, 22] indicate that $K_n$, proportional to $m$, is a good choice although $K_n$ should be a function of the dimensionality $N$ also.

Fukunanga and Hostetler [25] derive an expression for the optimum value of $K_n$ (in the mean-square-error sense) which is a function of $N$, $m$, and the true underlying density $p(x)$. If the underlying distribution is multivariate Gaussian, $K_{opt}$, the optimum value of $K_n$, is proportional to $m^{1/(N+1)}$. The constant of proportionality reduces from 0.75 for $N = 4$ to 0.42 for $N = 32$. Note that in one dimension ($N = 1$), $K_{opt}$ is substantially larger than the commonly used heuristic value of $m^{1/2}$. Also for a fixed sample size, as the dimensionality increases, $K_n$ goes down, which perhaps can be attributed to sparseness of the data. Vajda and Fritz [63] make some corrections, for small values of $N$, in the expression for $K_{opt}$ in [25]. Rozell and Kolvek [54] point out that $K_{opt}$ in the context of density estimation is also applicable to the $K$-nearest-neighbor decision rule. Thus the results of [25] and [63] can be used to select the optimum number of nearest neighbors for decision making.

4. Error estimation

The performance of a pattern classifier is usually evaluated by determining its classification accuracy. In most theoretical studies in pattern recognition one computes the expected recognition accuracy. In practice, since the true underlying distributions are not known, one estimates the recognition accuracy from the available samples. This generally involves dividing the total available samples into design and test sets; the classifier is designed based on the samples in the design set and is evaluated by estimating the error rate on samples in the test set [29, 42, 61].

Ideally one would like to use as many samples as possible in designing as well as in testing the classifier. However, it is well known that the design-set error rate (same set of samples are used to design as well as to test the classifier) is a biased estimate of the true error rate [61]. In fact, Cover [18] showed that for a two-class problem if a linear discriminant function is used, then the design-set error rate is always zero as long as $(m + m^2) < (N + 1)$. What then are the situations in which the design-set error rate can be 'safely' used as an estimate of the true error rate? Foley [24] considered this problem when the class-conditional densities are multivariate Gaussian with unknown mean vectors $\mu_i$ and $\mu_j$, and known covariance matrix $\Sigma$. He computed the expected design-set error rate as a function of $m$ (number of samples per class), $N$ and $\Sigma$, the true Mahalanobis distance. Foley's results can be summarized as follows:

1. $\lim_{m \to \infty} E(\epsilon_d(m, N, \Sigma) = 0) = 0$, where $E_d$ denotes the design-set error rate. That is, by adding more and more measurements it is possible to make the design-set error rate approach zero even if the true error rate is finite.

2. The ratio $(m/N)$ is critical to the bias in the design-set error rate. If $(m/N) > 3$, then the bias in the design-set error rate is close to zero.

Mehrotra [47] extended Foley's results to situations where the common covariance matrix $\Sigma$ is also unknown and concluded that the ratio $(m/N)$ must be larger than five before the bias in the design-set error rate is sufficiently small. These results seem to confirm the hypothesis [38, 39]: "the less is known about the underlying probability structure, the larger is the ratio of sample size to dimensionality".

Intuitively it appears that one should know the class assignment of test samples in order to estimate the error rate of a classifier. However, in many applications of pattern recognition methodology, labelling of samples can be very expensive. In view of this, Chow [16] proposed a procedure to estimate the error rate of a classifier based on a set of unlabelled test samples. Chow established a relationship between the error rate and the reject rate of a classifier, and since computing the reject rate does not require the knowledge of class assignment of test samples, the error rate can be determined with unlabelled test samples. Again, the ratio of the number of training samples to dimensionality plays an important role in this method of estimating the error rate. Fukunaga and Kessel [26] showed that for two Gaussian distributions with unknown mean vectors and unknown common covariance matrix, the estimate of error rate obtained from the empirical reject rate is optimistically biased. This bias in the error rate is a function of the ratio of sample size to dimensionality, and [26] recommends that this ratio must be at least ten for the bias to be small.

5. Conclusions

We have shown the important role which dimensionality and sample size play in various areas of pattern recognition, namely, classification accuracy, $K$-nearest-neighbor approach and error estimation. There is no doubt that the designer of a pattern recognition system should make every possible effort to obtain as many samples as possible. As the number of samples increases, not only does the designer have more confidence in the performance of the classifier, but more measurements can be incorporated in the design of the classifier without the fear of peaking in its performance. However, there are many pattern classification problems where either the number of samples is limited (for example, in a medical decision-making problem, there may be only a small number of patients available who are suffering from a specific disease) or obtaining a large number of samples
is extremely expensive (as in the application of pattern recognition methodology to nondestructive testing where specific defects in a piece of metal have to be detected by, say, analyzing the return ultrasonic waveform). It is in these small sample size problems where the designer of a recognition system has to be extremely careful. As we have pointed out in this paper, if the designer chooses to take the optimal Bayesian approach, then the average performance of the classifier improves monotonically as the number of measurements is increased. Most practical pattern recognition systems employ a non-Bayesian decision rule because the use of optimal Bayesian approach requires knowledge of prior densities, and besides, their complexity precludes the development of real-time recognition systems. The peaking behavior of practical classifiers is caused principally by their nonoptimal use of measurements.

Realizing the dangers of having too many measurements in the classifier design when the number of samples is small, a designer must select a ‘‘good’’ subset of measurements. Various techniques of feature selection and extraction [32, 38, 41] have been proposed in the literature for this purpose. However, most of these techniques use some sort of a criterion function such as a distance or information measure to select a good subset of measurements which must be estimated from the available samples. If the number of available samples is small compared to dimensionality, these estimates themselves will not be very reliable [31, 49, 71]. The same argument holds for finding the intrinsic dimensionality of a given set of data [51].

In summary, while the ratio of sample size to dimensionality is a crucial factor in the design of a recognition system, very little quantitative information is available to a designer. If the designer knows that the underlying distributions are multivariate Gaussian with unknown parameters, then some tables are available to determine the optimal number of measurements to use for a given sample size. However, since these tables are generated based on expected recognition accuracy, the designer must know the true Mahalanobis distance to use them. The general guideline for having five to ten times as many samples as measurements still seems to be a good practice to follow. We share our opinion expressed in [38, 39] that the ratio of sample size to dimensionality should be inversely proportional to the amount of knowledge about the class-conditioned densities.

References

[27] Galley, W. R. (1951). Discriminatory analysis: Perfect discrimination as the number of variables increases. Report No. 5, Project No. 21-49-004, USDA School of Aviation Medicine, Randolph Field, TX.


