Pergamon

Pattern Recognition, Vol. 29, No. 5, pp. 873–887, 1996
Elsevier Science Ltd
Copyright © 1996 Pattern Recognition Society
Printed in Great Britain. All rights reserved
0031–3203/96 $15.00 + .00

0031–3203(95)00122–0

A CLASS DISCRIMINABILITY MEASURE BASED ON FEATURE SPACE PARTITIONING

ANDRÉ F. KOHN,† LUÍS G. M. NAKANO‡ and MIGUEL OLIVEIRA E SILVA§

† Laboratório de Engenharia Biomédica, Departamento de Engenharia Eletrônica, Escola Politécnica, Universidade de São Paulo, Cx. P. 61548, CEP 05424-970, São Paulo, S. P., Brazil

‡ Laboratório de Sistemas Digitais, Departamento de Engenharia de Computação, Escola Politécnica, Universidade de São Paulo, Brazil

§ Departamento de Electrónica e Telecomunicações, Universidade de Aveiro, Portugal

(*Received* 29 *September* 1994; *in revised form* 4 *August* 1995; *received for publication* 15 *August* 1995)

Abstract—This paper presents a new class discriminability measure based on an adaptive partitioning of the feature space according to the available class samples. It is intended to be used as a criterion in a classifier-independent feature selection procedure. The partitioning is performed according to a binary splitting rule and appropriate stopping criteria. Results from several tests with Gaussian and non-Gaussian, multidimensional and multiclass computer-generated samples, were very similar to those obtained using a Bayes error criterion function, i.e. the optimal feature subsets selected by both criterion functions were the same. The main advantage of the new measure is that it is computationally efficient.

Class discriminability measure Feature selection criterion function Variable selection criterion Feature evaluation Interclass distance measure Class separability measure

1. INTRODUCTION

One of the first tasks in the design of a pattern classification system is to choose features (also called variables, measurements or attributes) that are thought to provide good discriminability among the existing pattern classes. This initial choice is based mainly on intuition and knowledge about the pattern generating mechanisms. If *all* the initially proposed features are used in the design and implementation of the pattern classification system, probably its performance will be inadequate, both in terms of misclassification rates and computation time and its hardware will be too expensive.

On the other hand, as the success or failure of a pattern recognition system is heavily dependent on the choice of good features (that separate well the classes), it is important not to limit *a priori* the number of features. Sometimes the choice of a new feature, never used before in a given application, may yield good performance to a classifier.[1,2] Therefore, initially, one should propose all the features that are judged potentially useful for the pattern classification system. Thereafter, a feature selection procedure should be used to choose a subset of the initial features according to quantitative criteria that will assure that the feature subset is among the best one could obtain from the original feature set. In feature selection the dimensionality is reduced by the elimination of features from an original feature set **H** of dimension n resulting in a feature subset **G** of dimension d ($d < n$). When the overall purpose is the design of a pattern classification system, the class samples projected on the features in subset **G** should be well discriminable.

A reduction in the number of features is necessary so that one may (1) decrease the cost (hardware and computer time) of measuring the features, (2) decrease the cost (hardware and computer time) of the classifier and (3) improve the performance (e.g. decrease the error rate) of the classifier.[3]

It should be added that for objectives (2) and (3), one may also use feature extraction which is the reduction of dimensionality obtained by a mapping from an n-dimensional feature-space to a d-dimensional space ($d < n$).[4,5]

The task of feature selection usually involves three choices: (a) a measure that quantifies the discriminability of the classes, (b) a search method to generate feature subsets from **H** and (c) a stopping criterion. An example from (b) would be a forward–backward search, also called "plus l–take away r",[6] and an example from (c) would be to stop at the best subset with $d = 12$ features when $n = 30$. Due to the focus of this paper, only item (a) will be discussed from now on. For overviews of many measures of class discriminability already described in the literature see references (6–8). The designer of the pattern classification system should choose carefully the measure of class discriminability to be employed in the feature selection procedure, since an inadequate choice will result in feature subsets exhibiting poor class discriminability and therefore the resulting pattern classifiers will have poor classification performances.

In some applications the type of classifier to be used in the pattern recognition system is predefined and in this case it may be desirable to search for good features using as a criterion the minimization of the classifier error rate. However, for each feature subset, one will have to design the classifier and then estimate the error rate, based on the available sample set by using, for example, the holdout method or another more elaborate technique such as the bootstrap.(9) Clearly, the error rate estimation approach to feature selection may lead to exorbitant computation times, particularly for high-dimensional data. Therefore, even when the classifier to be used is already defined, it may be interesting to have an independent feature selection stage, not based on error counting.

In other applications it may not be desired to choose the classifier in advance and hence the feature selection stage has to be carried out independently of the design of a classifier.

When the overall problem at hand is the design of a good classifier, class discriminability measures to be used in a feature selection procedure should preferably give an idea of classifier error rate. The Bayes classifier is theoretically the best that could be used and, therefore, an attractive discriminability measure is the Bayes error rate. Unfortunately, it is rather awkward to use in practice because its estimation by direct methods is computationally very intensive (also, the distribution of the classes is usually unknown and hence the Bayes classifier cannot be designed exactly, but only approximated by estimates based on the available class samples). There are probabilistic discriminability measures that provide an upper or lower bound for the Bayes error rate for the special case of two classes. Examples are the Jeffries–Matusita and the Bhattacharyya distance.(6,7)·However, the computational effort for this type of discriminability measure is quite considerable, due to the integrals of the (usually unknown) probability densities involved. An exception is the Gaussian case for which the distance expressions are simple, depending only on the mean vectors and covariance matrices. However, even for the Gaussian case it is not clear how to extend any of the known probabilistic discriminability measures to the multiclass case. One idea is to define a weighted average of any of the two-class measures over all pairs of classes; other suggestions may be found in references (10, 11). On the other hand, probabilistic dependence and information theoretic discriminability measures(10) have the advantage of being defined for an arbitrary number of classes. Some of these probabilistic discriminability measures provide bounds for the Bayes error rate,(10,11) but their practical usefulness is not clear both in terms of theoretical considerations (tightness of the bounds, effects of estimation errors) and practical implementation. For many, the computational effort is greater than the direct computation of Bayes error rate, while it is not clear in which multiclass cases the performance of these discriminability measures may parallel that of the Bayes error rate measure in a feature selection procedure. To decrease the computational effort, other class discriminability measures have been proposed, such as $\mathrm{tr}(\mathrm{W}^{-1}\mathrm{B})$, where W and B are the within-class and between-class scatter matrices, respectively.(6,7) For classes with different covariance matrices, W is an average within-class scatter matrix. In any case, these simpler discriminability measures are even less related to the Bayes error rate than those measures based on the full probabilistic class description. They are potentially very useful for practical applications due to their computational simplicity, but are expected to work properly only when the distributions are unimodal with separated sample means.(7)

The purpose of this paper is to present a class discriminability measure for arbitrary distributions and any number of classes and that is fast to compute. There is no *a priori* knowledge required about the classes, the measure being based on the available set of random samples from each class. Therefore, the measure proposed is directly oriented towards practical applications.

The proposed class discriminability measure is based on a data-adaptive partitioning of the feature space in such a way that regions having samples from two or more classes are more finely partitioned than the ones with samples from a single class. After the partitioning is over, there will be homogeneous and nonhomogeneous "buckets", i.e. approximately hypercubic subregions with samples from a single class and from multiple classes, respectively. The class discriminability measure proposed is based on the samples in the nonhomogeneous buckets. Several tests run with Gaussian and non-Gaussian, multidimensional and multiclass synthetic data indicate that its performance parallels that of the criterion of minimizing the Bayes error rate.

The basic concept of the partitioning technique employed here is similar to that used in the construction of decision tree classifiers,(12,13) but the partitioning obtained here is specific for the objective of this work which is to find a (computer efficient) class discriminability measure and not to design a particular classifier. Once the partition of the feature space is achieved, a class discriminability measure is defined based on the samples from regions of the partitioned space that have samples from more than one class.

2. DESCRIPTION OF THE ALGORITHM TO COMPUTE THE CLASS DISCRIMINABILITY MEASURE

The algorithm to compute the class discriminability measure proposed in this paper may be subdivided in three stages, the first two dealing with the construction of the feature space partition and the third dealing with the computation of the class discriminability index itself.

Let S be a set containing N vector samples $\mathbf{x}$ $(n \times 1)$ obtained from c classes. The relative frequency of samples from any class is an estimate of the *a priori* class probability. Each vector $\mathbf{x}$ of n features representing a given pattern has a label indicating to which of the c classes it belongs.

2.1. *Feature-space partitioning: splitting a subregion in two others*

Initially, the ranges of the samples from S in the direction of all n features are determined and a hyper-rectangular parallelepiped is constructed such that its faces are perpendicular to the feature axes and located at the maxima and minima of the samples in each feature direction. For example, let $x_{1_{\min}}$ and $x_{1_{\max}}$ be the minimum and maximum values obtained from the projections of all the samples in S on the direction of feature x_1. Therefore, the above mentioned hyper-rectangular parallelepiped will have two faces on the two parallel hyperplanes orthogonal to axis x_1, passing through points $x_{1_{\min}}$ and $x_{1_{\max}}$. When this is completed for all feature directions, the hyper-rectangular parallelepiped faces are defined by the intersection of all the hyperplanes. To make the wording shorter, we shall term any hyper-rectangular parallelepiped a *box*.

A given box at any stage of the partitioning of the feature-space is first tested for the stopping conditions presented below. Assuming none of the stopping criteria are satisfied, the box is partitioned into two other boxes. The split is performed along the feature that has the largest range (in the samples). The starting point for the splitting is at a position corresponding to the median of the samples projected on that feature coordinate. The median generates two sample subsets around which the two new boxes are built. This splitting rule is almost the same as that proposed by reference (12), the difference being that in our method the boundaries are always associated with samples. The feature-space partitioning is finished when there are no remaining boxes to be split, i.e. when all boxes, also called *terminal boxes*, satisfy at least one of the stopping criteria listed below.

2.2. *Feature-space partitioning: stopping criteria*

The following stopping criteria were developed for the feature-space partitioning algorithm, having in mind that the purpose is that of finding an index of class discriminability:

(1) if the box is homogeneous, i.e. it contains samples from a single class,

(2) if the samples in the box are from linearly separable classes,

(3) if the number of samples in the box is less than N^a, where N is the total number of samples in S and $a = 0.375$.

If *any* criterion is satisfied then the box under analysis is not split any more.

The *first* stopping criterion is rather obvious, since our objective is to somehow quantify the degree of overlap of the samples from different classes. If seen from a Bayes classifier viewpoint, a homogeneous box containing samples only from class w_i indicates a region where the (estimate of the) *a posteriori* class probability is equal to one and hence that region does not contribute to the Bayes error rate.

The *second* criterion was developed to cover regions where the samples from two or more classes are not overlapping. These regions would not contribute to the Bayes error rate. The usefulness of this second criterion is two-fold. In a case where the classes are linearly separable with a hyperplane that is not orthogonal to any of the feature axes, there would be many box splittings until a stopping criterion could be satisfied. This is due to the fact that any splitting in the algorithm can only be carried out along a coordinate axis. With this second criterion many of these unnecessary box splittings would be avoided. Another case is when the classes are separable with nonlinear hypersurfaces. Here, this criterion is expected to be useful in smaller boxes, where the corresponding sections of the nonlinear boundaries may be approximately linear. In order to avoid increasing too much the computational effort, a simplified test for linear discriminability was chosen, described in the following. Assuming there are samples from b classes in a given box, the corresponding b centroids are found and all the lines joining these centroids are formed. The following test is carried out for each line: samples in the box are projected onto the line and if the projections of samples from each class do not overlap with any of the projections of the other classes, then the condition of linear discriminability along that line is satisfied. The overall linear discriminability in the box is considered satisfied if it is satisfied for all lines. As this is a simplified test, some cases of linear discriminability will not be detected and in some applications the user may want to apply a more rigorous test of linear discriminability.

The *third* stopping criteron is needed to avoid the occurrence of very small boxes (e.g. having one, two or three samples) at the end of the space partitioning. Very small or very large boxes would tend to cause an increase in the bias of the class discriminability measure defined in section *C* below. Very large boxes would give a large bias, because they usually do not carry a fine enough representation of the region where there is some overlap between the class samples. At the end of the feature-space partition procedure one is left with a collection of boxes of different sizes and contents. If the samples in each box should be able to give a reasonable estimate of the *a posteriori* class probabilities in the region within each box (see section *C* below), or in other words, the class-conditional probability density functions, then one can view this problem as rather similar to the probability density estimation by the k-nearest neighbor method (k-NN). Here the main difficulty is the determination of the value of k. Fukunaga and Hostetler[14] derived the optimal value of k to be used in k-NN density estimation but the expression is too complicated to be used for general distributions. Fukunaga and Hummels[15] point out the difficulties of using any of the available theoretical values for k when the objective is to obtain estimates of the Bayes error rate. Since our class discriminability measure is expected to have a relation with the Bayes error rate, we are left with the equivalent difficulty of choosing the maximum number of samples for a nonhomogeneous not linearly-separable terminal box. Fukunaga and Hummels[15] show that to select a good value for k it is useful to run many experiments with different values for k and observing the behavior of the corresponding Bayes error rate estimates. Due to these difficulties in the selection of a value for k, we chose somewhat arbitrarily the value N^a, with $a = 3/8$, for the maximum number of samples for a nonhomogeneous nonseparable terminal box. This relation was suggested by Enas and Choi[16] as being a good value for k in the design of some k-NN classifiers. Also, this value for k is seen to give good results when one analyses the experimental results of Fukunaga and Hummels[15] for data sets of dimensions 8 and 60. We think this upper limit for the number of samples in a nonhomogeneous terminal box will work adequately for many practical cases. Perhaps in some specific situations more refined relations for k could be employed, including a dependence also on the dimension of the space and on the subregion being analysed, as suggested by the theoretical work of Fukunaga and Hostetler.[14]

Summarizing, after the end of the partitioning procedure one is left with:

(i) homogeneous terminal boxes (HTB),

(ii) nonhomogeneous linearly separable terminal boxes (NLSTB) and

(iii) nonhomogeneous not linearly separable terminal boxes (NNLSTB)

2.3. *Computation of the class discriminability measure*

The class discriminability measure (CDM) is based on the M nonhomogeneous, not linearly-separable terminal boxes (NNLSTBs) resulting from the space partitioning procedure described in subsections 2.1 and 2.2. It is defined as:

$$\mathrm{CDM} = \frac{1}{N}\sum_{i=1}^{M}\{k(i) - \max_{j}[k(j|i)]\},$$

where $k(i)$ is the total number of samples in the ith NNLSTB, $k(j|i)$ is the number of samples from class j in the ith NNLSTB and N is the overall number of samples.

This measure was motivated by a discrete approximation to the Bayes error rate E^* as seen in the following:

$$E^* = \int_{\Omega}\left\{1 - \max_{j} P(w_j|\mathbf{x})\right\} p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

where $P(w_j|\mathbf{x})$ is the *a posteriori* probability of class $w_j, j = 1,\dots,c$ and Ω is the feature space. In approximation:

(a) only the NNLSTB will contribute to the estimate of E^*;

(b) $P(w_j|\mathbf{x}) \cong k(j|i)/k(i)$ for $\mathbf{x} \in i$th NNLSTB;

(c) $\mathrm{d}\mathbf{x} \cong \Delta V(\mathbf{x})$;

and therefore:

$$E^* \cong \sum_{i=1}^{M}\left\{1 - \max_{j}[k(j|i)/k(i)]\right\} \cdot [k(i)/(N\cdot\Delta V(\mathbf{x}))]\cdot\Delta V(\mathbf{x})$$

and finally

$$E^* \cong \frac{1}{N}\sum_{i=1}^{M}\left\{k(i) - \max_{j}[k(j|i)]\right\} = \mathrm{CDM}.$$

In terms of the algorithm, for each nonhomogeneous not linearly separable terminal box one subtracts the number of samples of the most frequent class from the total number of samples in that box. All these values are summed and the result is divided by N. Actually, as it is difficult to quantify the influence of the proposed adaptive feature space partitioning algorithm in terms of the expression for CDM, this index should be seen as potentially useful for the task of searching for good features in a pattern recognition problem. Perhaps different details in the space partitioning algorithm could lead to closer estimates of the Bayes error rate, but that is not the main objective of the present paper.

3. METHODS

The objective of this work was to find a computationally efficient class discriminability measure (CDM) that could perform well for problems with arbitrary distributions. Therefore, it seemed interesting to compare the performance of our CDM with the estimates of the Bayes error rate and with the trace of $W^{-1}B$, the latter being a popular class discriminability index which is fast to compute. As a theoretical approach did not seem feasible, all the comparisons were based on several computer-simulated data sets covering both Gaussian and non-Gaussian cases.

3.1. *Generation of Gaussian sets of samples*

Multivariate Gaussian (G) sample vectors, with given mean vectors and covariance matrices, were obtained from Gaussian univariate samples by means of standard techniques. Several different cases were generated: samples with "reasonable" overlap between classes (RO), samples with "little" overlap between classes (LO), samples with dimension 4 ($4D$), samples with dimension 5 ($5D$), samples with dimension 9 ($9D$), samples for 2 classes ($2C$), samples for 3 classes ($3C$), samples for equal *a priori* class probabilities (EP) and finally, samples for different *a priori* class probabilities (DP). For one specific Gaussian test, 10 sets were generated from the same distributions to enable the analysis of an average behavior of the three class discriminability measures, this being indicated as ($10x$). For future reference, each set of samples is coded using the abbreviations above. For example, a set of Gaussian samples with reasonable overlap, feature-space dimension equal to 4 and having samples from 3 classes with different *a priori* probabilities would have the code *G/RO/4D/3C/DP*. A few details from each sample set is presented in the following, without the inclusion of the corresponding mean vectors and covariance matrices due to space limitations. The covariance matrices were designed so that the axes of the hyper-ellipsoids were not orthogonal to the coordinate axes and were not parallel for the different classes involved. The only difference between the sample sets types *EP* and *DP* were the number of samples per class; the mean vectors and covariance matrices being the same.

Sample set G/LO/4D/2C/EP: there are 1000 samples per class.

Sample set G/LO/4D/2C/DP: there are 1000 samples for class 1 and 3000 samples for class 2, i.e. the desired *a priori* class probabilities are 0.25 and 0.75.

Sample set G/LO/4D/3C/EP: there are 1000 samples per class.

Sample set G/LO/4D/3C/DP: there are 1000 samples for class 1, 1500 samples for class 2 and 500 samples for class 3, i.e. the *a priori* class probabilities are 0.3333, 0.5 and 0.1667, respectively.

Sample set G/RO/4D/2C/EP: there are 1000 samples for each of the two classes.

Sample set G/RO/4D/2C/DP. there are 1000 samples for class 1 and 3000 samples for class 2.

Sample set G/RO/4D/3C/EP: there are 1000 samples for each of the three classes.

Sample set G/RO/4D/3C/CP: there are 1000 samples for class 1, 1500 for class 2 and 500 for class 3.

Sample sets G/RO/4D/3C/DP/10*x*: 10 different sample sets were generated, each having 1000, 1500 and 500 samples from classes 1, 2 and 3, respectively, following the same distributions used for the previous sample set described above. For this purpose, 1900, 2400 and 1400 samples were initially generated from classes 1, 2 and 3, respectively; the first set was formed from the first 1000, 1500 and 500 samples from classes 1, 2 and 3. The second set was formed by ignoring the first 100 samples for each class and adding the next new 100 samples for each class, and so on. This method was chosen for storage economy and to simulate a more realistic situation of limited availability of samples.

Sample sets G/LO/9D/2C/EP and *G/RO/9D/2C/EP*: there are 1000 samples per class, each sample being a nine-dimensional vector.

Sample sets G/LO/9D/2C/DP and *G/RO/9D/2C/DP*: there are 1000 samples from class 1 and 3000 samples from class 2.

Sample sets G/LO/9D/3C/EP and *G/RO/9D/3C/EP*: there are 1000 samples per class.

Sample sets G/LO/9D/3C/DP and *G/RO/9D/3C/DP*: there are 1000 samples from class 1, 1500 from class 2 and 500 from class 3.

Sample set G/5D/2C/EP/LVR: as a final set of samples, we generated a two-class case in which there was a large variance ratio (LVR) for the two classes. The space dimensionality was equal to 5, the features were statistically independent and the variances of class 1 over features 1 to 4 was 3.7×10^{-1} while the corresponding variances for class 2 were 3.7×10^{-2}. Along feature 5 the variances were equal to 10^{-1} for both classes. The mean for class 1 was at the origin and for class 2 was at $[0.50\,0.75\,1.00\,1.25 \text{ and } 0.50]^T$. The number of samples per class was 1000.

3.2. *Generation of non-Gaussian sets of samples*

Multivariate non-Gaussian (NG) sample vectors were generated as follows: along features 1 and 2 the samples were Gaussian for class 1 and non-Gaussian for class 2, while along the remaining features they were jointly Gaussian. Features 1 and 2 were statistically independent from the other features. The non-Gaussian samples (from class 2) in the direction of features 1 and 2 were generated from a mixture of r Gaussians, resulting in a multimodal distribution. For example, in Fig. 1 the bordering samples are from class 2 and are formed by five subclouds, each associated with a different Gaussian distribution. In terms of covariance matrices, for the ith subcloud from class 2 there is an associated overall $(n \times n)$ block-diagonal covariance matrix Ci_2, having a full 2×2 and a full $(n-2) \times (n-2)$ submatrices in the main diagonal, with the remaining terms being zero (representing the statistical independence of the first two features from the others). Therefore, there are r $(n \times n)$ covariance matrices Ci_2 and r mean vectors mi_2 $(i = 1, \ldots, r)$ defining the overall statistics of class 2 in n-dimensional space, while for class 1 a single $(n \times n)$ covariance matrix C_1 and a single $(n \times 1)$ mean vector m_1 define the overall statistics, Two different non-Gaussian cases were examined: one which we term a partial surround (PS), exemplified in Fig. 1, and another we term a total surround (TS), exemplified in Fig. 2. A brief summary of the non-Gaussian sample sets generated is given in the following.

Sample set NG/PS/5D/2C/EP: this is a set containing five-dimensional samples from two classes, being 1000 per class. Samples along features 1 and 2 associated with class 1 come from a bivariate Gaussian distribution and can be seen in Fig. 1 as the central elliptic distribution, centered at the origin. The bordering cloud (hence the name Partial Surround—PS) in Fig. 1 is from class 2 and is formed by a mixture of five different bivariate Gaussian subclouds, each with 200 samples. Class 1 and 2 samples along the other features are from Gaussians, but independent from those along features 1 and 2.

Sample sets NG/PS/5D/2C/EP/10*x*: 10 different sample sets were generated, each with 1000 samples from each class, following the distributions used in the previous sample set above. A similar procedure to that already described for *G/RO/4D/3C/DP*/10*x* was employed here, where each of the ten sample sets came from a superset with 1900 samples from class 1 and 380

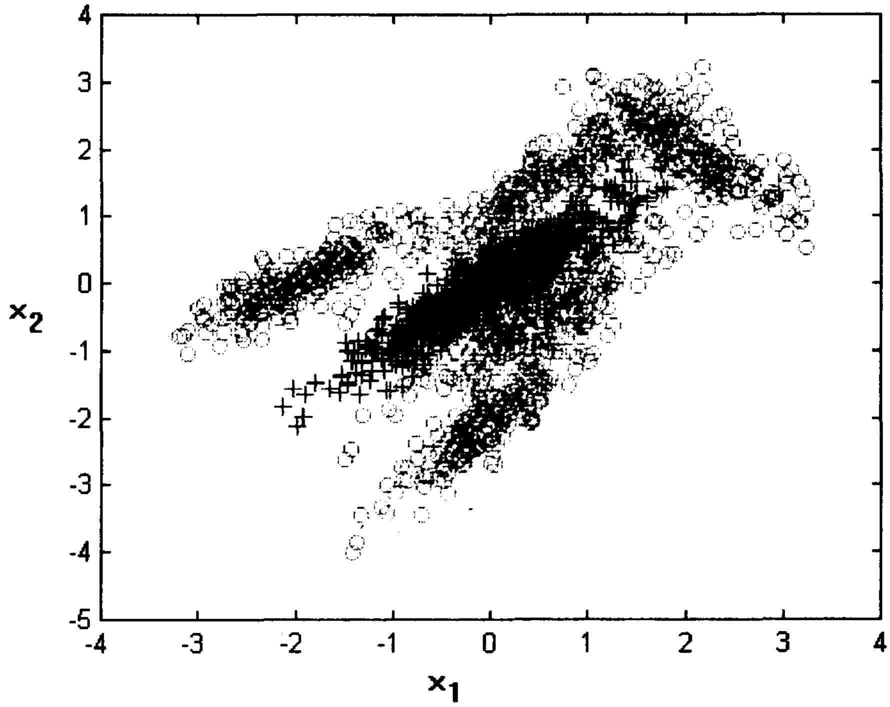

Fig. 1. Samples from set *NG/PS/5D/2C/EP* projected on feature coordinates x_1 and x_2.

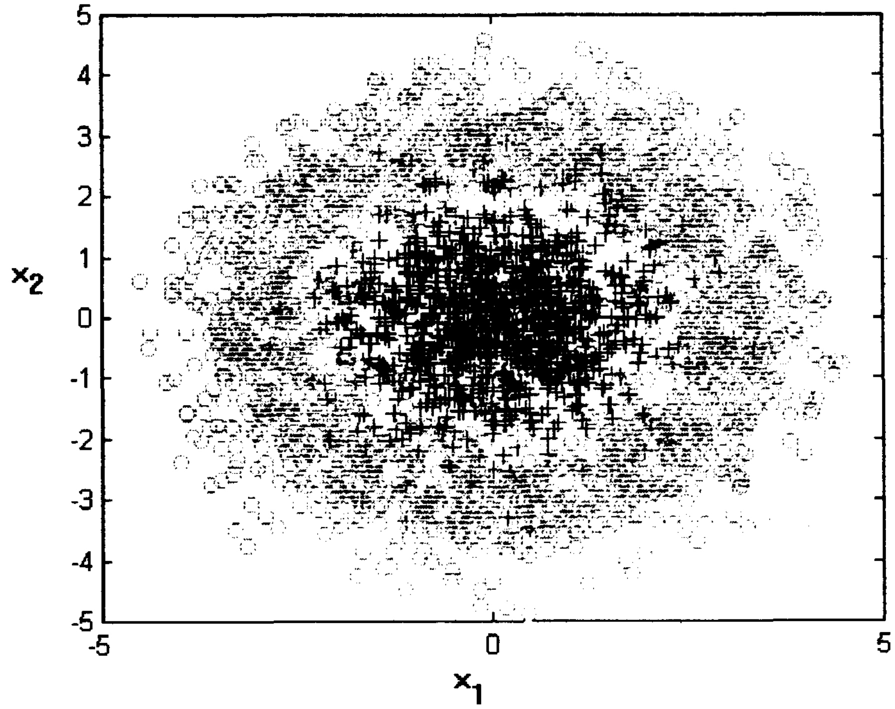

Fig. 2. Samples from set *NG/TS/5D/2C/DP* projected on feature coordinates x_1 and x_2.

samples from each of the five Gaussians associated with class 2.

Sample set NG/TS/5D/2C/DP: there are 1000 samples from class 1 and 1600 samples from class 2. Along features 1 and 2, the samples from class 1 originate from a single Gaussian with mean vector at the origin (see Fig. 2). Along the same coordinates, the surrounding annular cloud (hence the name *T*otal *S*urround—TS) of samples in Fig. 2 arises from eight different Gaussian distributions, with 200 samples from each.

The samples along the remaining coordinates were generated independently from those along features 1 and 2.

3.3. *Evaluation of the new class discriminability measure*

For each set of samples S a feature selection procedure was run either for a preselected number of features d ($d < n$) or for all numbers of features, from $d = 1$ to $d = n$ (see below). For any given d-dimensional feature subset, the following class discriminability measures were estimated from the data in S: the Bayes error rate, the trace of $W^{-1}B$ and that presented in this work, based on *f*eature *s*pace *p*artitioning (FSP measure, for short). An exhaustive analysis of all d-dimensional feature subsets was carried out for each d and a ranking of the subsets was obtained according to the criterion values, the best subset being associated with the lowest values either for the Bayes or FSP measures and associated with the highest value for the trace measure. In two cases, 10 different sample sets were generated from the same distributions and the most frequent subset was found (more or less in the spirit of Murray(17)) besides the average values for each class discriminability measures.

The Bayes error rate was estimated by counting the errors of the optimal Bayes decision rule in classifying the data in S.

The computation of $\mathrm{tr}(W^{-1}B)$ from the available samples was carried out in a straightforward way. The corresponding matrices W and B for a given feature subset were obtained from the n-dimensional versions by adequate selection of rows and columns. Both the Bayes and the trace criteria algorithms were implemented in C++ (GCC compiler).

The feature-space-partitioning measure was implemented according to the algorithm described in the previous section using C programming language and compiler generating tools (Bison and Flex, from Free Software Foundation). All the other programs were implemented in C++ language (GCC compiler), with standard libraries (libg++) and were run mostly in Silicon Graphics Power Series 480 VGX (with eight R3000 processors) and on Sun SparcStation IPC computers.

4. EXPERIMENTAL RESULTS

This section will present the results corresponding to the several sample sets described in the previous section. The three class discriminability measures to be compared are the Bayes error rate, the trace of $W^{-1}B$ and that based on *f*eature *s*pace *p*artitioning (FSP). Whenever needed, the three measures will be denoted by the short forms Bayes, trace and FSP.

4.1. *Results obtained with Gaussian sets of samples*

For *sample set G/LO/4D/2C/EP*, the feature selection procedure was run only for $d = 2$, i.e. the objective was to study the class discriminability measures for all the subsets with two features when the available number of features was $n = 4$. The three measures selected the pair of features $\{1, 4\}$ as the one that gave the best class discriminability. Table 1 shows all the feature pairs and the corresponding values for the Bayes, trace and FSP measures. The ordering of the features was the same for all the three measures.

For *sample set G/LO/4D/2C/DP*, the three measures again gave feature pair $\{1, 4\}$ as the best. Table 2 shows that the first four feature pairs were equally ordered in the three measures, but there was an inversion between $\{2, 3\}$ and $\{2, 4\}$ for the trace and FSP measures.

For *sample set G/LO/4D/3C/EP*, the best feature pair was $\{1, 2\}$ for all three measures. As a matter of

Table 1. Feature subsets and the values of the three class discriminability measures for the *sample set G/LO/4D/2C/EP*

Rank	Subset—Bayes		Subset—Trace		Subset—FSP	
1	1, 4	0.0015	1, 4	7.180074	1, 4	0.0165
2	1, 2	0.0075	1, 2	5.674424	1, 2	0.0290
3	1, 3	0.0235	1, 3	3.445907	1, 3	0.0950
4	3, 4	0.0465	3, 4	2.727266	3, 4	0.1810
5	2, 3	0.0490	2, 3	2.120405	2, 3	0.2430
6	2, 4	0.0765	2, 4	2.044600	2, 4	0.2815

Table 2. Feature subsets and the values of the three class discriminability measures for the *sample set G/LO/4D/2C/DP*

Rank	Subset—Bayes		Subset—Trace		Subset—FSP	
1	1, 4	0.00075	1, 4	9.529667	1, 4	0.00575
2	1, 2	0.00550	1, 2	6.720351	1, 2	0.01575
3	1, 3	0.02250	1, 3	3.479530	1, 3	0.07000
4	3, 4	0.03450	3, 4	2.621752	3, 4	0.10725
5	2, 3	0.04600	2, 4	1.918607•	2, 4	0.15575
6	2, 4	0.05475	2, 3	1.737918	2, 3	0.17800

Table 3. Feature subsets and the values of the three class discriminability measures for the *sample set G/LO/4D/3C/EP*

Rank	Subset—Bayes		Subset—Trace		Subset—FSP	
1	1, 2	0.01767	1, 2	13.63981	1, 2	0.05567
2	1, 4	0.02533	1, 4	12.84676	1, 4	0.07000
3	1, 3	0.03400	1, 3	8.618266	1, 3	0.10933
4	2, 3	0.05633	2, 3	5.536205	2, 3	0.19533
5	2, 4	0.09700	3, 4	4.852386	2, 4	0.30300
6	3, 4	0.10233	2, 4	4.728407	3, 4	0.31000

Table 4. Feature subsets and the values of the three class discriminability measures for the *sample set G/LO/4D/3C/DP*

Rank	Subset—Bayes		Subset—Trace		Subset—FSP	
1	1, 2	0.0213	1, 4	16.08054	1, 2	0.0413
2	1, 4	0.0277	1, 2	14.76135	1, 4	0.0457
3	1, 3	0.0397	1, 3	8.536114	1, 3	0.1063
4	2, 3	0.0757	3, 4	5.116483	2, 3	0.2443
5	2, 4	0.1120	2, 3	4.703231	3, 4	0.2510
6	3, 4	0.1127	2, 4	4.370039	2, 4	0.2623

Table 5. Feature subsets and the values of the three class discriminability measures for the *sample set G/RO/4D/2C/EP*

Rank	Subset—Bayes		Subset—Trace		Subset—FSP	
1	1, 2	0.1295	1, 2	1.277910	1, 2	0.4075
2	1, 3	0.1655	1, 3	0.972301	1, 3	0.4150
3	1, 4	0.1800	1, 4	0.907602	2, 3	0.4615
4	2, 3	0.1805	2, 3	0.801263	1, 4	0.4900
5	2, 4	0.1915	2, 4	0.720001	2, 4	0.5445
6	3, 4	0.2425	3, 4	0.481692	3, 4	0.5705

fact, all three measures gave the same ordering of feature pairs, as shown in Table 3.

For *sample set G/LO/4D/3C/DP*, the best feature subset was again {1, 2} for both the Bayes and FSP, but the trace criterion chose the feature subset {1, 4}. Nevertheless, it should be emphasized that the Bayes ranking for these two feature pairs was due to a single sample and therefore both pairs for this sample set are practically equivalent. The trace criterion chose the pair {1, 2} as the second best among all the feature pairs. Table 4 shows that the rankings given by both the Bayes and FSP measures coincide up to the four best pairs. The fourth pair for the trace criterion is {3, 4}, which is the worst pair according to the Bayes criterion, the difference between it and pair {2, 3} is associated with 111 misclassified samples. The difference between pairs {2, 4} and {3, 4} is associated with only two misclassified samples.

For *sample set G/RO/4D/2C/EP*, where the samples from the two classes were more overlapped than in the first set described in this section, the best feature pair was {1, 2} for all the three criteria, as shown in Table 5. Pairs {1, 4} and {2, 3} differ in terms of the Bayes criterion by a single misclassified sample and therefore are practically equivalent.

For *sample set G/RO/4D/2C/DP*, the same ordering of feature pairs was achieved by all three criteria (from best to worst): {1, 2}, {1, 3}, {1, 4}, {2, 3}, {2, 4} and {3, 4}. The corresponding sequence of Bayes error rate estimates were: 0.10700, 0.12650, 0.13825, 0.15600, 0.16875 and 0.17525.

For *sample set G/RO/4D/3C/EP*, where the samples from three classes had a reasonable overlap, the best feature pair for all three criteria was {1, 2}. Table 6 shows that for the third best feature pairs both the trace and the FSP criteria pointed to pair {1, 4} instead of the pair {2, 3} indicated by the Bayes criterion. This does not seem bad because the difference between both pairs is only six misclassified samples, while the difference between pairs {1, 2} and {1, 3} is 117 misclassified samples.

For *sample set G/RO/4D/3C/DP*, the three criteria selected as the best feature pair {1, 2}. Table 7 shows the feature subset orderings for $d = 1$, 2 and 3. For $d = 1$ and $d = 3$, all three criteria gave the same subset ordering. For $d = 2$, the trace criterion gave a similar feature ordering than the Bayes criterion. The FSP criterion inverted the second and third best feature pairs, which in terms of the Bayes error estimates differ by 37 misclassified samples. It is interesting to note

Table 6. Feature subsets and the values of the three class discriminability measures for the *sample set G/RO/4D/3C/EP*

Rank	Subset—Bayes		Subset—Trace		Subset—FSP	
1	1, 2	0.10667	1, 2	5.881687	1, 2	0.29400
2	1, 3	0.14567	1, 3	4.770069	1, 3	0.34800
3	2, 3	0.16233	1, 4	4.061874	1, 4	0.39000
4	1, 4	0.16433	2, 3	3.771261	2, 3	0.40000
5	2, 4	0.17933	2, 4	3.210390	2, 4	0.42467
6	3, 4	0.26833	3, 4	1.991239	3, 4	0.47233

Table 7. Feature subsets and the values of the three class discriminability measures for the *sample set G/RO/4D/3C/DP*

Rank	Subset—Bayes		Subset—Trace		Subset—FSP	
$d=1$						
1	1	0.22000	1	2.982694	1	0.44667
2	2	0.23533	2	2.300813	2	0.47433
3	3	0.32133	3	1.168982	3	0.51633
4	4	0.37600	4	0.734204	4	0.55000
$d=2$						
1	1, 2	0.12667	1, 2	5.323766	1, 2	0.27467
2	1, 3	0.16000	1, 3	4.306004	1, 4	0.36667
3	1, 4	0.17233	1, 4	3.701928	1, 3	0.37100
4	2, 3	0.18567	2, 3	3.226132	2, 3	0.42000
5	2, 4	0.19567	2, 4	2.822457	2, 4	0.43533
6	3, 4	0.26867	3, 4	1.880330	3, 4	0.46033
$d=3$						
1	1, 2, 3	0.10133	1, 2, 3	6.385880	1, 2, 3	0.25467
2	1, 2, 4	0.10700	1, 2, 4	5.830955	1, 2, 4	0.28100
3	1, 3, 4	0.13567	1, 3, 4	5.000775	1, 3, 4	0.35067
4	2, 3, 4	0.16600	2, 3, 4	3.740582	2, 3, 4	0.39733
$d=4$						
1	1, 2, 3, 4	0.09133	1, 2, 3, 4	6.885136	1, 2, 3, 4	0.23900

that on passing from the best pair to the second, the Bayes index increased by 26.3%, while the FSP increased by 33.5%. On the other hand on passing from the second to the third best respective feature pairs, the Bayes index increased by 7.7% while the FSP increased by 1.2% i.e. both criteria indicated clearly the almost equivalence of pairs {1, 3} and {1, 4}. Table 7 also shows that the three class discriminability measures showed a monotonic behavior for increasing numbers of features when for each d the corresponding best subset was taken. Another interesting observation is that, in this example, the best feature subsets for each value of d are nested.

For *sample sets G/RO/4D/3C/DP/*10*x*, the average values for each of the class discriminability measures were obtained from the 10 sample sets. The same qualitative results were obtained as those shown in Table 7 (e.g. the same feature subset sequence for each criterion, nesting of features) and, therefore, will not be presented. Also, for each d and for each criterion, the best ranked feature subsets were the same for each of the 10 individual sample sets. These results suggest that the findings are robust.

For *sample set G/LO/9D/2C/EP*, the objective was to find the best subset with $d=4$ features from the original set of $n=9$ features. The results are synthesized in Table 8, where the 10 best feature subsets are indicated for each of the three criteria. The three criteria indicated the same feature subset {1, 5, 7, 9} as being the best among all four-tuples of features. It should be added that, at least for the Bayes and FSP criteria, this choice was robust in the sense that the percentage increase of the Bayes and FSP measures from the first to the second feature subsets was large. As to the 10 best subsets indicated by the Bayes criterion, seven were also indicated by the FSP criterion and nine by the trace criterion (but in different ordering) among their respective 10 best subsets.

For *sample set G/LO/9D/2C/DP*, again the objective was to select the best subset with $d=4$ features. Table 9 shows that both the Bayes and the FSP criterion indicated the subset {1, 5, 7, 9} as the best while this same subset was ranked fifth by the trace criterion. The best subset for the trace criterion was ranked sixth by the Bayes criterion. Here it is clear that the trace criterion did not work properly, while the FSP did. As

Table 8. Feature subsets and the values of the three class discriminability measures for the *sample set G/LO/9D/2C/EP*

Rank	Subset—Bayes		Subset—Trace		Subset—FSP	
1	1, 5, 7, 9	0.0040	1, 5, 7, 9	6.22778	1, 5, 7, 9	0.0080
2	1, 3, 5, 9	0.0055	1, 2, 7, 9	5.850402	1, 3, 7, 9	0.0290
3	3, 5, 7, 9	0.0065	1, 3, 5, 9	5.821537	1, 2, 4, 7	0.0420
4	1, 3, 7, 9	0.0075	1, 4, 5, 9	5.518858	1, 3, 4, 9	0.0425
5	1, 5, 6, 9	0.0080	3, 5, 7, 9	5.117251	1, 3, 4, 7	0.0435
6	1, 4, 5, 9	0.0095	1, 4, 7, 9	4.979232	1, 3, 5, 9	0.0520
7	1, 4, 7, 9	0.0095	3, 4, 7, 9	4.913938	1, 4, 7, 9	0.0550
8	3, 4, 7, 9	0.0095	1, 2, 4, 5	4.747062	1, 2, 4, 5	0.0560
9	1, 2, 4, 7	0.0110	1, 3, 4, 9	4.654403	1, 5, 6, 9	0.0625
10	1, 2, 4, 5	0.0115	1, 2, 4, 7	4.578616	1, 2, 5, 9	0.0650

Table 9. Feature subsets and the values of the three class discriminability measures for the *sample set G/LO/9D/2C/DP*

Rank	Subset—Bayes		Subset—Trace		Subset—FSP	
1	1, 5, 7, 9	0.00325	1, 4, 5, 9	6.451560	1, 5, 7, 9	0.02300
2	1, 3, 7, 9	0.00500	1, 3, 5, 9	6.441597	1, 4, 5, 9	0.02950
3	1, 3, 5, 9	0.00525	1, 3, 7, 9	6.005819	1, 5, 6, 9	0.03275
4	3, 5, 7, 9	0.00575	3, 4, 7, 9	5.978488	1, 3, 5, 9	0.03350
5	3, 4, 7, 9	0.00675	1, 5, 7, 9	5.619694	1, 3, 7, 9	0.03350
6	1, 4, 5, 9	0.00750	3, 5, 7, 9	5.551622	4, 5, 7, 9	0.03625
7	1, 2, 4, 7	0.00875	1, 3, 4, 9	5.001707	3, 5, 7, 9	0.03725
8	1, 4, 7, 9	0.00875	3, 6, 7, 9	4.896725	1, 3, 4, 9	0.03800
9	1, 3, 4, 7	0.00925	3, 7, 8, 9	4.876851	1, 2, 5, 9	0.04175
10	1, 4, 5, 7	0.00950	2, 3, 7, 9	4.860629	1, 2, 4, 7	0.04250

Table 10. Feature subsets and the values of the three class discriminability measures for the *sample set G/LO/9D/3C/EP*

Rank	Subset—Bayes		Subset—Trace		Subset—FSP	
1	1, 2, 4, 5	0.01633	1, 2, 4, 5	10.34657	1, 2, 4, 5	0.06200
2	1, 2, 5, 6	0.01767	1, 5, 7, 9	10.12569	1, 2, 5, 9	0.08100
3	1, 2, 5, 9	0.02233	1, 2, 5, 9	10.03391	1, 2, 5, 6	0.08433
4	2, 4, 5, 8	0.03233	1, 2, 5, 7	8.795321	2, 3, 5, 9	0.09700
5	2, 4, 5, 7	0.03267	1, 2, 5, 6	8.767937	2, 4, 5, 8	0.10067
6	1, 2, 3, 5	0.03400	1, 2, 3, 5	8.626267	2, 4, 5, 9	0.10467
7	1, 2, 5, 7	0.03433	1, 2, 5, 8	8.305842	2, 4, 5, 7	0.10633
8	2, 4, 5, 6	0.03433	1, 4, 5, 9	8.055298	1, 4, 5, 9	0.11533
9	2, 4, 5, 9	0.03467	1, 3, 7, 9	7.667932	2, 4, 5, 6	0.11700
10	2, 3, 4, 5	0.03500	2, 5, 7, 9	7.574672	1, 2, 5, 7	0.11900

to the 10 best subsets indicated by the Bayes criterion, six were also indicated by the FSP and trace criteria (with different ranking) among their best 10 subsets.

For *sample set G/LO/9D/3C/EP*, the three criteria indicated the same subset {1, 2, 4, 5} as the best among all feature subsets with cardinality 4. Table 10 shows that among the best five subsets indicated by the Bayes criterion four (three) were also indicated by the FSP (trace) criterion. Among the 10 best Bayes subsets the FSP indicated eight and the trace only five. In this test the subset {1, 2, 4, 5} (chosen by all three criteria as the best) was only slightly better (four samples difference) than subset {1, 2, 5, 6} according to the Bayes criterion.

For *sample set G/LO/9D/3C/DP*, the Bayes and FSP criteria indicated {1, 2, 4, 5} as the best subset with four features, while the trace criterion indicated the subset {1, 5, 7, 9}. Surprisingly, the latter is not included among the best 10 subsets indicated by the Bayes criterion.

For *sample set G/RO/9D/2C/EP*, where the classes were a little more overlapped, Table 11 shows that the three criteria indicated the same optimal subset {1, 2, 4, 6}. Among the best five subsets indicated by the trace and the FSP criteria, three subsets are included among the first five indicated by the Bayes criterion. Seven (six) subsets among the best 10 indicated by the trace (FSP) criteria are included among the best subsets indicated by the Bayes criterion. Again in this test the best quadruplet {1, 2, 4, 6} is only a few samples (5) better than the next according to the Bayes criterion.

Table 11. Feature subsets and the values of the three class discriminability measures for the *sample set G/RO/9D/2C/EP*

Rank	Subset—Bayes		Subset—Trace		Subset—FSP	
1	1, 2, 4, 6	0.0395	1, 2, 4, 6	3.238273	1, 2, 4, 6	0.1600
2	1, 2, 3, 4	0.0420	1, 2, 3, 4	2.897928	1, 2, 3, 5	0.1710
3	1, 2, 4, 9	0.0495	1, 3, 4, 6	2.844463	1, 2, 3, 4	0.1715
4	1, 3, 4, 6	0.0505	2, 3, 4, 6	2.671908	1, 2, 4, 9	0.1715
5	1, 3, 4, 8	0.0505	1, 2, 3, 6	2.652638	1, 2, 4, 5	0.1745
6	2, 4, 6, 7	0.0505	2, 4, 6, 7	2.645448	1, 4, 5, 6	0.1810
7	2, 3, 4, 6	0.0515	1, 2, 4, 8	2.616946	1, 4, 6, 9	0.1860
8	1, 2, 4, 5	0.0520	1, 3, 4, 8	2.616763	1, 3, 4, 9	0.1925
9	2, 4, 6, 9	0.0530	2, 4, 6, 8	2.587327	1, 2, 4, 7	0.1930
10	1, 3, 4, 9	0.0545	2, 4, 6, 9	2.566807	1, 3, 4, 8	0.1960

Table 12. Feature subsets and the values of the three class discriminability measures for the *sample set G/RO/9D/2C/DP*

Rank	Subset—Bayes		Subset—Trace		Subset—FSP	
1	1, 2, 4, 6	0.03125	1, 2, 4, 6	3.265063	1, 2, 4, 6	0.13675
2	2, 4, 6, 7	0.03950	2, 4, 6, 7	2.837815	1, 2, 4, 9	0.13875
3	1, 2, 3, 4	0.04075	2, 4, 6, 8	2.753324	2, 4, 6, 9	0.13875
4	2, 3, 4, 6	0.04225	1, 2, 3, 4	2.681162	2, 4, 6, 7	0.14125
5	2, 4, 6, 8	0.04325	2, 3, 4, 6	2.679840	2, 4, 6, 8	0.14200
6	1, 3, 4, 6	0.04350	2, 4, 6, 9	2.677656	1, 2, 3, 4	0.14700
7	2, 4, 5, 6	0.04425	2, 4, 5, 6	2.619490	1, 2, 5, 6	0.14775
8	2, 4, 6, 9	0.04525	1, 2, 4, 8	2.598846	1, 2, 4, 5	0.14925
9	1, 2, 4, 9	0.04575	1, 2, 4, 9	2.567216	1, 2, 6, 7	0.14975
10	1, 2, 4, 5	0.04650	1, 2, 4, 5	2.557409	1, 2, 4, 7	0.15025

For *sample set G/RO/9D/2C/DP*, the three criteria selected the feature subset {1, 2, 4, 6} as the best among all subsets with four features. Table 12 shows that five (three) subsets indicated by the trace (FSP) criterion among its first five were also indicated by the Bayes criterion, although with a different ranking. Among the best 10 subsets indicated by the trace (FSP) criterion, nine (seven) also belong to the list of the 10 best subsets indicated by the Bayes criterion.

For *sample set G/RO/9D/3C/EP*, the subset {1, 2, 4, 6} was indicated as the best subset with four features by the three criteria. Among the best 10 subsets indicated by the Bayes criterion, the trace (FSP) criterion indicated eight (seven) among its list of 10 best subsets.

For *sample set G/RO/9D/3C/DP*, again subset {1, 2, 4, 6} was indicated by the three criteria as being the best among all subsets with cardinality four. Five (three) among the first five best subsets according to the trace (FSP) criterion were among the best five subsets according to the Bayes criterion. Seven (six) among the best 10 subsets indicated by the trace (FSP) criterion are among the best 10 indicated by the Bayes criterion.

For *sample set G/5D/2C/EP/LVR*, the best pair of features for the Bayes and the FSP criteria was {3, 4}, with values 0.0145 and 0.0270. The second and third best subsets for the Bayes were {2, 4} and {1, 4}, with values 0.0205 and 0.0230, i.e. differing between them by only five misclassified samples. These same two subsets were the third and second best, respectively, for the FSP criterion, with values 0.0495 and 0.0470.

4.2. *Results obtained with non-Gaussian sets of samples*

For *sample set NG/PS/5D/2C/EP*, the two classes were non-Gaussian along features 1 and 2, as seen in Fig. 1, with samples from class 2 partially surrounding those from class 1. Along other coordinates the distributions were Gaussian. The feature selection procedure was run for all numbers of features *d*, using the three class discriminability measures. Table 13 summarizes the findings.

For the single feature search the Bayes criterion indicated features {3} and {4} as being the best and second best, respectively, while the FSP criterion chose them in reverse order, i.e. feature {4} as being the best. Nevertheless, as features {3} and {4} differ in terms of the Bayes index by only 16 misclassified samples out of 2000 while along feature {3} there were 422 misclassified samples, it seems that the two features are practically equivalent in terms of class discriminability. The trace criterion indicated the correct best feature {3}, but indicated feature {5} as the second best.

Table 13 shows the results for the search of the pairs of features, indicating that the best subset according to the Bayes criterion, {1, 2}, was also the best according to the FSP criterion, but it was the worst pair according to the trace criterion. The estimates of the Bayes

Table 13. Feature subsets and the values of the three class discriminability measures for the *sample set NG/PS/5D/2C/EP*

Rank	Subset—Bayes		Subset—Trace		Subset—FSP	
$d = 1$						
1	3	0.2110	3	0.500938	4	0.6205
2	4	0.2190	5	0.489283	3	0.6775
3	5	0.2290	4	0.056039	5	0.6835
4	2	0.2590	1	0.011594	1	0.6870
5	1	0.2935	2	0.008052	2	0.6910
$d = 2$						
1	1, 2	0.0270	3, 4	1.048146	1, 2	0.1490
2	3, 4	0.0720	4, 5	0.979414	3, 4	0.3275
3	4, 5	0.0865	1, 5	0.739464	4, 5	0.4030
4	1, 4	0.1250	1, 3	0.730588	1, 4	0.5020
5	2, 3	0.1325	3, 5	0.622067	1, 3	0.5115
6	1, 5	0.1370	2, 3	0.605914	1, 5	0.5140
7	2, 4	0.1395	2, 5	0.588242	2, 3	0.5345
8	2, 5	0.1395	1, 4	0.057185	2, 5	0.5590
9	1, 3	0.1475	2, 4	0.056042	2, 4	0.5880
10	3, 5	0.1865	1, 2	0.012952	3, 5	0.6400
$d = 3$						
1	1, 2, 4	0.0115	3, 4, 5	1.393755	1, 2, 4	0.1170
2	1, 2, 3	0.0200	1, 3, 4	1.214799	1, 2, 5	0.1475
3	1, 2, 5	0.0220	1, 4, 5	1.167876	1, 2, 3	0.1705
4	2, 3, 4	0.0365	2, 3, 4	1.064226	1, 4, 5	0.1735
5	2, 4, 5	0.0390	2, 4, 5	0.994330	2, 4, 5	0.1965
6	1, 4, 5	0.0475	1, 3, 5	0.988816	1, 3, 4	0.2050
7	1, 3, 4	0.0495	2, 3, 5	0.768057	2, 3, 4	0.2110
8	3, 4, 5	0.0640	1, 2, 5	0.746557	3, 4, 5	0.3305
9	2, 3, 5	0.1160	1, 2, 3	0.742059	1, 3, 5	0.4145
10	1, 3, 5	0.1170	1, 2, 4	0.057614	2, 3, 5	0.4695
$d = 4$						
1	1, 2, 3, 4	0.0055	1, 3, 4, 5	1.743554	1, 2, 4, 5	0.0800
2	1, 2, 4, 5	0.0075	2, 3, 4, 5	1.424391	1, 2, 3, 4	0.1125
3	1, 2, 3, 5	0.0170	1, 2, 3, 4	1.220483	2, 3, 4, 5	0.1660
4	2, 3, 4, 5	0.0275	1, 2, 4, 5	1.177042	1, 2, 3, 5	0.1665
5	1, 3, 4, 5	0.0330	1, 2, 3, 5	1.002338	1, 3, 4, 5	0.1730
$d = 5$						
1	1, 2, 3, 4, 5	0.0050	1, 2, 3, 4, 5	1.757253	1, 2, 3, 4, 5	0.0980

error rates for the pairs {1, 2} and {3, 4} were 2.70 and 7.20%, the difference being associated with 90 misclassified samples, suggesting that the choice of the pair {1, 2} is robust. Figure 1 shows the samples projected on these coordinates. Pair {3, 4} was selected by the trace criterion as the best among all feature pairs. Finally, among the best five pairs of features indicated by the FSP (trace) criterion, four (two) were among the first five pairs ranked by the Bayes criterion.

The search for the best triplet of features resulted in the selection of subset {1, 2, 4} by both the Bayes and FSP criteria, while the trace criterion selected subset {3, 4, 5}, which is the eighth ranked by the Bayes index. Among the five best ranked feature subsets for the FSP (trace) criterion, four (two) were among the five best indicated by the Bayes criterion.

The feature selection procedure using the Bayes criterion indicated subset {1, 2, 3, 4} as the best among the five possible combinations of four features. The FSP criterion indicated the subset {1, 2, 4, 5}, which is the second best subset according to the Bayes error measure, being associated with only four additional misclassified samples when compared with the Bayes best subset {1, 2, 3, 4}. The second-ranked subset according to the FSP criterion is the best indicated by the Bayes. On the other hand, the trace criterion selected the subset {1, 3, 4, 5}, which is the worst according to the Bayes criterion, being associated with 61 additional misclassified samples when compared to the subset {1, 2, 3, 4}.

Two additional observations can be made on the results of Table 13. One is that for the Bayes and FSP criteria there was no nesting of attributes when d was increased from 1 to 5. This is an important observation in terms of the choice of feature selection search procedures.[6] Another aspect is that for the Bayes and trace criteria there was monotonicity in the respective class discriminability measure values when d was increased from 1 to 5 and the best subset was chosen for each d. For the FSP measure there was a (slight) break in the monotonicity on passing from three to four and then to five features (0.1170 to 0.0800 to 0.098). This

Table 14. Feature subsets and the values of the three class discriminability measures for the *sample set NG/PS/5D/2C/EP/*10*x*

Rank	Subset—Bayes		Subset—Trace		Subset—FSP	
$d = 1$						
1	4	0.2123	3	0.478699	4	0.6498
2	3	0.2181	5	0.450200	3	0.6775
3	5	0.2345	4	0.060381	1	0.6895
4	2	0.2506	1	0.008814	5	0.6900
5	1	0.2957	2	0.005379	2	0.6930
$d = 2$						
1	1, 2	0.0291	3, 4	1.019153	1, 2	0.1470
2	3, 4	0.0711	4, 5	0.936467	3, 4	0.3250
3	4, 5	0.0897	1, 3	0.639107	4, 5	0.4065
4	1, 4	0.1207	1, 5	0.615236	1, 4	0.4680
5	2, 3	0.1332	3, 5	0.583945	1, 5	0.5125
6	2, 4	0.1380	2, 3	0.576764	2, 3	0.5185
7	2, 5	0.1401	2, 5	0.537873	1, 3	0.5370
8	1, 5	0.1432	2, 4	0.060594	2, 4	0.5470
9	1, 3	0.1534	1, 4	0.055081	2, 5	0.5475
10	3, 5	0.1894	1, 2	0.009559	3, 5	0.6345
$d = 3$						
1	1, 2, 4	0.0122	3, 4, 5	1.342712	1, 2, 4	0.1105
2	1, 2, 3	0.0212	1, 3, 4	1.195138	1, 2, 5	0.1430
3	1, 2, 5	0.0226	1, 4, 5	1.128206	1, 4, 5	0.1570
4	2, 3, 4	0.0334	2, 3, 4	1.037295	1, 2, 3	0.1650
5	2, 4, 5	0.0387	2, 4, 5	0.950745	2, 3, 4	0.1760
6	1, 3, 4	0.0445	1, 3, 5	0.932925	2, 4, 5	0.1800
7	1, 4, 5	0.0486	2, 3, 5	0.719692	1, 3, 4	0.2105
8	3, 4, 5	0.0655	1, 2, 3	0.709955	3, 4, 5	0.3315
9	2, 3, 5	0.1111	1, 2, 5	0.686433	1, 3, 5	0.4420
10	1, 3, 5	0.1198	1, 2, 4	0.061812	2, 3, 5	0.4470
$d = 4$						
1	1, 2, 3, 4	0.0065	1, 3, 4, 5	1.713666	1, 2, 4, 5	0.0925
2	1, 2, 4, 5	0.0077	2, 3, 4, 5	1.377449	1, 2, 3, 4	0.1040
3	1, 2, 3, 5	0.0175	1, 2, 3, 4	1.199954	1, 3, 4, 5	0.1445
4	2, 3, 4, 5	0.0252	1, 2, 4, 5	1.134706	2, 3, 4, 5	0.1535
5	1, 3, 4, 5	0.0325	1, 2, 3, 5	0.946600	1, 2, 3, 5	0.1645
$d = 5$						
1	1, 2, 3, 4, 5	0.0042	1, 2, 3, 4, 5	1.725232	1, 2, 3, 4, 5	0.0920

does not seem worrisome because the Bayes index for the case of four features was 0.0055 and that for five features was 0.0050, i.e. a difference of a single misclassified sample. In other words, in this example it will not make too much of a difference if you use four or five features in your classifier.

For *sample sets NG/PS/5D/2C/EP/*10*x*, the average values for the three class discriminability measures were computed from the 10 sample sets. Table 14 shows that the results were reasonably similar to those found for a single sample set (Table 13), even though there were a few different rankings of feature subsets. Another difference was that the small loss of monotonicity for the FSP found in the previous case (see Table 13) did not happen here. The values of the three criteria varied monotonically for the best feature subsets when d varied from 1 to 5 (even though the FSP values for the best subset of four features and for the set of five features were practically equal). Finally, for *each* of the ten sample sets (not shown), the same optimal subsets were always obtained for each d and each criterion. For example, for criterion FSP, the optimal subset with two features was found to be $\{1, 2\}$ for each of the 10 sample sets. These results suggest that the findings are robust.

For *sample set NG/TS/5D/2C/DP*, along features $\{1, 2\}$ the samples from class 2 totally surrounded those from sample 1, as seen in Fig. 2. The feature selection procedure was run only for $d = 2$, with Table 15 summarizing the results obtained. Both the Bayes and the FSP criteria indicated $\{1, 2\}$ as the best pair among the 10 possible pairs of features. On the other hand the trace criterion indicated $\{4, 5\}$ as the best, which was ranked as the third by Bayes and with a quite higher error rate estimate (13.3% compared with 7.3%). The best four pairs selected by the FSP were among the five best from the Bayes criterion.

A short remark should be made on the selection of the parameter a in the expression N^a, used in the third stopping criterion of the feature space partitioning procedure. The results presented here all used $a = 0.375$, which showed a good feature selection

Table 15. Feature subsets and the values of the three class discriminability measures for the *sample set NG/TS/5D/2C/DP*

Rank	Subset—Bayes		Subset—Trace		Subset—FSP	
1	1, 2	0.07308	4, 5	0.028801	1, 2	0.14462
2	1, 5	0.12846	1, 5	0.021860	1, 5	0.31923
3	4, 5	0.13269	3, 5	0.021114	2, 5	0.33654
4	2, 5	0.14077	2, 5	0.020007	4, 5	0.34115
5	2, 4	0.15115	3, 4	0.000503	3, 5	0.37846
6	3, 5	0.16308	1, 4	0.000468	2, 4	0.38038
7	1, 4	0.16692	2, 4	0.000465	1, 4	0.42500
8	2, 3	0.22154	1, 3	0.000030	2, 3	0.47385
9	3, 4	0.22846	2, 3	0.000026	1, 3	0.50115
10	1, 3	0.23192	1, 2	0.000024	3, 4	0.50962

performance for the data sets used in the present work. Smaller values for a tended to give FSP measures closer to the estimated Bayes error rates, but too small values (e.g. 0.200) sometimes made the optimal feature subset choice to be in error. A value which tended to give good results, and better approximations to the Bayes error rate, was $a = 0.250$, except for the cases where the superposition of the classes was small.

4.3. *A comparison of computation times*

Processing times for the evaluation of all feature pairs for the non-Gaussian, total surround, five-dimensional feature space sample set (*NG/TS/5D/2C/DP*) were obtained for three computers: an IBM-PC 486-compatible (33 MHz, 8 Mb RAM), a Sun SPARC station IPC with 12 Mb RAM and a Silicon Graphics Power Series 480 VGX (eight processors R3000, 256 Mb total memory, 8 kb of cache memory for instructions per processor, 1 Mb of total cache per processor). The same compiler was used in all machines (GCC 2.5.7 from Free Software Foundation). None of the programs developed to compute the three criteria were optimized with respect to computation time. Table 16 shows the results obtained, where it can be concluded that the FSP measure was about two times slower than the trace measure, but at least 13 times faster than the Bayes measure. Nevertheless, it must be emphasized that the Bayes measure employed here assumed all the distributions known *a priori* and hence each classifier could be designed *a priori*. In a more practical case one would have to estimate probability density functions from the available data samples before counting the errors to estimate the error rate. Obviously in this more realistic setting the processing times for the Bayes measure would be prohibitive. As the trace and FSP were implemented based totally on the available samples, the corresponding computation times are of practical relevance.

Table 16. Computation times for the three class discriminability measures for the case *NG/TS/5D/2C/DP*

	Silicon Graphics	Sun SPARC station	IBM-PC
Trace	4.3 s	28.3 s	3 min
FSP	7.8 s	66.2 s	5 min
Bayes	5 min	15 min	N.A.

5. CONCLUSION

This paper proposes a class discriminability measure defined by sample counts on the partitioned feature space (FSP measure). Its performance in a feature selection procedure was compared with that of the estimated Bayes error rate and also the trace of $W^{-1}B$. The former was used as the "golden standard" for the evaluation of the feature selection *per se* and the latter was used as a reference for the computational efficiency.

Several tests (more than 20 are presented here) were run with controlled artificial data in an effort to cover many different and relevant situations. The analyses of the results showed that the FSP measure behaved practically as well as the Bayes index, but with a much higher computation efficiency.

The main results and conclusions of this work are:

(a) For practically all experimental tests employed in this work the *optimal* feature subsets indicated by the FSP measure were the same as those given by the Bayes index (rank 1 in all the tables). On the other hand, the trace of $W^{-1}B$ gave very poor results for the non-Gaussian cases analysed and was slightly inferior to the FSP for the Gaussian cases.

(b) For two cases (Gaussian and non-Gaussian data sets) the robustness of the FSP-based method was confirmed by running the feature selection procedure for 10 sample sets obtained from the same probability density functions.

(c) One test was run to give an idea of the resolution of the FSP measure. The variance of one class was 10 times that of the other and the FSP index was also able to indicate the optimal feature subset.

(d) The computational efficiency of the feature selection procedure with the FSP measure was at least an order of magnitude better than with the Bayes measure and only about twice worse than with the trace measure. It is important to emphasize that the Bayes computation times were "small" because we used *a priori* knowledge about the class distributions and hence the optimal classifier was determined independent of the data samples. On the other hand, if nonparametric estimates of the Bayes measure (or other measure involving probability density functions) based on, e.g. kernel or nearest-neighbors would be employed in a practical feature selection problem, the involved computational times would probably be prohibitive, many orders of magnitude larger than those obtained here. It should be emphasized that with a fast and good performance class discriminability measure an exhaustive feature subset search could be feasible in many real life problems thereby avoiding the pitfalls of suboptimal search strategies.

There are many suggestions one could make in an effort to improve the algorithms, both in terms of processing time and class discriminability performance. For example, α-trimmed means could be used instead of medians; the choice of the next feature for box partitioning could use class information and not only spread of the samples; a more refined test for linear separability could be employed; the third stopping criterion in the space partitioning could depend not only on the total number of samples, but also on the space dimensionality and perhaps on local conditions.

Acknowledgements—This work was supported in part by CNPq and Capes (Brasil) and JNICT (Portugal). The authors are thankful to Carlos Minoru Muto, from Escola Politécnica da USP, for his help and for his valuable comments. We also thank LSI (Escola Politécnica da USP) for computer time in the Silicon Graphics.

REFERENCES

1. A. F. Kohn, Fractal number and spectral skewness: two features for the pattern classification of motor unit action potentials, *Proc. 11th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 1885–1886. Seattle, Washington, U.S.A. (November 1989).
2. H. Bakardjian, Ventricular beat classifier using fractal number clustering, *Med. Biol. Eng. Comput.* **30**, 495–502 (1992).
3. A. K. Jain and B. Chandrasekaran, Dimensionality and sample size considerations in pattern recognition practice, *Handbook of Statistics*, P. R. Krishnaiah and L. N. Kanal, eds, Vol. 2, pp. 835–855. North-Holland, Amsterdam (1982).
4. C. Lee and D. A. Landgrebe, Feature extraction based on decision boundaries, *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-15**, 388–400 (1993).
5. Y. Hamamoto, T. Kanaoka and S. Tomita, On a theoretical comparison between the orthonormal discriminant vector method and discriminant analysis, *Pattern Recognition* **26**, 1863–1867 (1993).
6. P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach.* Prentice-Hall, New Jersey (1982).
7. K. Fukunaga, *Statistical Pattern Recognition.* Academic Press, San Diego (1990).
8. G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition.* Wiley, New York (1992).
9. B. Efron, Bootstrap methods: another look at the jackknife, *Ann. Statist.* **7**, 1–26 (1979).
10. M. Ben-Bessat, Use of distance measures, information measures and error bounds in feature evaluation, *Handbook of Statistics*, P. R. Krishnaiah and L. N. Kanal, eds, Vol. 2, pp. 773–791. North-Holland, Amsterdam (1982).
11. N. Glick, Separation and probability of correct classification among two or more distributions, *Ann. Inst. Statist. Math.* **25**, 373–382 (1973).
12. G. R. Dattatreya and L. N. Kanal, Decision trees in pattern recognition, *Progress in Pattern Recognition 2*, L. N. Kanal and A. Rosenfeld, eds, pp. 189–239. North-Holland, Amsterdam (1985).
13. L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, *Classification and Regression Trees.* Wadsworth, Belmont (1984).
14. K. Fukunaga and L. D. Hostetler, Optimization of k-nearest-neighbor density estimates, *IEEE Trans. Inform. Theory* **IT-19**, 320–326 (1973).
15. K. Fukunaga and D. M. Hummels, Bayes error estimation using Parzen and k-NN procedures, *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-9**, 634–643 (1987).
16. G. G. Enas and S. C. Choi, Choice of the smoothing parameter and efficiency of k-nearest neighbor classification, *Comput. Maths. Appl.* **12A**, 235–244 (1986).
17. G. Murray, A cautionary note on selection of variables in discriminant analysis, *Appl. Statist.* **26**, 246–250 (1977).

About the Author—ANDRÉ FABIO KOHN received his B.S. and M.S. in electrical engineering from the University of São Paulo, Escola Politécnica (Brasil) and his Ph.D. in Engineering from the University of California at Los Angeles, in 1973, 1976 and 1980, respectively. He is a Full Professor at the University of São Paulo, Escola Politécnica, where he teaches and researches signal processing, pattern recognition and experimental and theoretical neurophysiology.

About the Author—LUÍS GUSTAVO MENDONÇA NAKANO received a B.S. degree in electrical engineering and an M.S. degree in computer engineering from the University of São Paulo, Escola Politécnica (Brasil), in 1991 and 1994, respectively. He is now a Ph.D. student at the University of Virginia. His fields of interest are pattern recognition and computer theory.

About the Author—MIGUEL OLIVEIRA E SILVA received his B.S. and M.S. degrees in electrical engineering from the University of Aveiro (Portugal) in 1990 and 1994, respectively. He is now a Ph.D. student at this same institution. His current research interests include software engineering, object-oriented concurrent programming, distributed systems, signal processing and pattern recognition.