10-1-1973

An Iterative Approach to the Feature Selection Problem

Henry P. Decell
University of Houston

John A. Quirein
TRW Systems Group

Conference on

Machine Processing of

Remotely Sensed Data

October 16 - 18, 1973

The Laboratory for Applications of
Remote Sensing

Purdue University
West Lafayette
Indiana

AN ITERATIVE APPROACH TO THE FEATURE SELECTION PROBLEM

Henry P. Decell, Jr.

Mathematics Department
University of Houston
Houston, Texas*

John A. Quirein

TRW Systems Group
Houston, Texas 77058

ABSTRACT

The B-average divergence for m-distinct classes, resulting from the linear transformation $y = Bx$, is proposed as a feature selection criterion, where B is a k by n matrix of rank $k \leq n$. It is shown that if the B-average divergence resulting from B is large enough, then the probability of misclassification, considered as a function of the class of all k by n matrices, is essentially minimized by B. A computer program, utilizing a gradient procedure, is developed to numerically maximize the B-average divergence and results are presented for the C1 flight line. For this example, corresponding to 9-distinct classes, most of the discriminatory information is found to lie in a 3-dimensional subspace, defined by an appropriately chosen 3 by 12 matrix B.

1. INTRODUCTION

This paper considers the problem of feature selection or reducing the dimension of the data to be processed from n to k. By reducing the dimension of the data from n to k, classification time is generally reduced. Yet the dimension reduction should not be so great that classification accuracy is impaired. Thus, consider the general problem of classifying an n-dimensional observation vector x into one of m-distinct classes π_i, $i=1,2,\ldots,m$ where each class π_i is normally distributed with mean μ_i and covariance Λ_i, so that we write $\pi_i = \pi_i(\mu_i, \Lambda_i)$. It can be shown (Anderson, 1958) that the probability of misclassification is minimized if a maximum likelihood classification procedure is used to classify the data. Thus, the notation PMC is used to denote this minimal probability of misclassification. The dimension of each observation vector to be processed can be conveniently reduced by performing the transformation $y = Bx$, where B is a k by n matrix of rank k. Thus, the n-dimensional classification problem transforms into a k-dimensional classification problem. The problem becomes one of classifying each k-dimensional observation vector y into one of m-distinct classes π_i, where now $\pi_i = \pi_i(B\mu_i, B\Lambda_i B^T)$. In this k-dimensional space determined by the row vectors of B, the minimal probability of misclassification resulting from applying a maximum likelihood classification procedure is denoted by PMC_B. Since the transformation $y = Bx$ produces a linear combination of the components of the observation vector x, it can be shown that, in general, information is lost and

$$PMC_B \geq PMC$$

Thus, for a fixed k, the feature selection problem could be stated as: select a k x n matrix $\hat{B}$ from the class of all k by n matrices of rank k such that

$$PMC_{\hat{B}} = \min PMC_B$$

*Work sponsored in part by the National Aeronautics and Space Administration, Johnson Space Center, Earth Observation Division, under Contract NAS 9-12777.

where PMC_B represents the probability of misclassification resulting from applying a maximum likelihood classification procedure on the transformed data Bx.

The problem of evaluating and minimizing PMC is handled indirectly. Let D(i,j) denote the interclass divergence between classes i and j (Kullback, 1968), as determined using n-dimensional information. Similarly, let $D_B(i,j)$ represent the interclass divergence between classes i and j resulting from performing the transformation y = Bx. It is noted that the interclass divergence is a measure of the "degree of difficulty" of discriminating between classes π_i and π_j with in general, the larger the interclass divergence, the greater the "separation" between classes π_i and π_j. Since (Kullback, 1968) it is true that

$$D(i,j) \geq D_B(i,j)$$

it follows that the difference

$$D(i,j) - D_B(i,j) \geq 0$$

can be considered as a measure of the separation to be gained for classes π_i and π_j. If the average divergence for m classes is defined by

$$D = c \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} D(i,j) \quad \text{where } c = \frac{2}{m(m-1)}$$

it follows that the "B-average divergence", D_B, satisfies

$$D_B = c \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} D_B(i,j) \leq c \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} D(i,j) = D$$

i.e., that $D_B \leq D$ for every k x n matrix B; k = 1,...,n.

We will prove the following theorem.

Theorem: If $D = D_B$, then $PMC_B = PMC$.

These results suggest for fixed k less than n, that one should select B so as to maximize D_B. Tou and Heydorn (1967) proposed a procedure to maximize $D_B(i,j)$, as a function of B. However, this procedure is valid only in case m = 2, i.e., the two-class problem. Babu (1972) extended the above procedure to the multi-class problem by proposing a procedure for maximizing D_B. Both procedures amount to computing the gradient of the appropriate function D_B or $D_B(i,j)$ with respect to B. Babu's expression for the gradient of the average divergence D_B with respect to B is (in addition to being incorrect) rather lengthy and numerically unattractive since it is expressed in terms of many eigenvalues and eigenvectors. In this paper, we derive a simple expression for the gradient of D_B with respect to B. This expression for the gradient is free of any requirement for computation of eigenvectors or eigenvalues, and, in addition, all matrix inversions necessary to evaluate the gradient are available from computing D_B. Thus, the feature selection problem becomes one of developing a numerical procedure to maximize D_B over the class of all k by n matrices of rank k. We will show that the maximum value of D_B is attained on the compact set, $\beta = [B:BB^T = I_k]$ and, further, that the maximum value of D_B is attained on $[B\epsilon\beta: B = (I_k|0)U$ where U is an element of the group of orthogonal transformations on $E^n]$.

The problem of selecting the "best" k is handled by obtaining the "best" B for various values of k less than n. Then an "adequate" value of k is determined by computing the difference $D - D_B$, and comparing D(i,j) with $D_B(i,j)$ for all distinct class pairs, where now, B is assumed to maximize D_B for a fixed k. The comparison of D(i,j) with $D_B(i,j)$ for all distinct class pairs constitutes what is called a "Class Separability to be Gained Map". For a given set of classes π_i and π_j, the value of $D_B(i,j)$ can be considered to represent the separability between classes π_i and π_j resulting from the transformation y = Bx. The difference $D(i,j) - D_B(i,j) \geq 0$ represents the separation to be gained for this class pair. Thus, we desire to find an integer k (preferably as small as possible) and corresponding optimal B such that the difference $D(i,j) - D_B(i,j)$ is "small" for all distinct class pairs.

A computer program, based on the mathematical results of the next section, was subsequently developed by TRW Systems to maximize D_B for a given k (Quirein, 1972). The program utilizes (in the iterative solution of the variational equation for B) a version (Johnson, 1969) of the Davidon Iterator (based on the Davidon-Fletcher-Powell technique), generously provided by Johnson Space Center. Numerical results for a 12-dimensional data set corresponding to 9 distinct classes obtained from Cl flight line data (Bond, 1972) are discussed in the final section.

2. MATHEMATICAL DEVELOPMENT

This section is included to derive and interpret the mathematical equations necessary to maximize the average divergence D_B numerically, and to relate the average divergence to the probability of misclassification. Also presented are additional mathematical results, obtained as an outgrowth of the University of Houston Mathematics Department Seminars in Pattern Recognition and Classification Theory. Many of the results derived below involve computing the partial derivative of a scaler ψ with respect to a matrix $B = \{b_{ij}\}$. We use the notation

$$\frac{\partial \psi}{\partial B}$$

to represent the matrix

$$\left(\frac{\partial \psi}{\partial b_{ij}}\right)$$

evaluated at B. Let

B; k by n matrix of rank $k \leq n$

Λ; n by n real symmetric matrix of rank n

S; n by n real symmetric matrix

and define

$$\psi = 1/2 \text{ tr } \{(B \Lambda B^T)^{-1} (B S B^T)\}$$

where tr denotes the trace of a matrix and superscript T denotes the transpose of a matrix. We prove the following Lemma.

Lemma 1

$$\left(\frac{\partial \psi}{\partial B}\right)^T = [SB^T - \Lambda B^T (B \Lambda B^T)^{-1} (B S B^T)](B \Lambda B^T)^{-1}$$

and thus

$$B\left(\frac{\partial \psi}{\partial B}\right)^T = (0)$$

Proof: Making use of the elementary properties of the trace of a matrix, and assuming no variation in the matrices Λ and S, it follows that the differential of ψ is given by

$$d\psi = \text{tr } \{dB[SB^T - \Lambda B^T (B \Lambda B^T)^{-1} (B S B^T)](B \Lambda B^T)^{-1}\}.$$

Thus, the result follows by noting

$$\frac{\partial \psi}{\partial b_{ij}} = \text{tr } \{\frac{\partial B}{\partial b_{ij}} [SB^T - \Lambda B^T (B \Lambda B^T)^{-1} (B S B^T)] (B \Lambda B^T)^{-1}\}.$$

It should be noted that any B-matrix corresponding to the selection of the transpose of any k-distinct eigenvectors of $\Lambda^{-1}S$ satisfies $\left(\frac{\partial \psi}{\partial B}\right)^T = (0)$. Lemma 1 shows that each row vector of $\frac{\partial \psi}{\partial B}$ is orthogonal to the subspace determined by the row vectors of B, so that we may assume in this case that $BB^T = I_k$, where I_k is a k by k identity matrix. Define the scaler

$$\Gamma = 1/2 \log |B\Lambda B^T|$$

where $\log |B\Lambda B^T|$ denotes the logarithm to the base e of the determinant of the matrix $B \Lambda B^T$.

Lemma 2

$$\left(\frac{\partial \Gamma}{\partial B}\right)^T = \Lambda B^T (B \Lambda B^T)^{-1}$$

and thus

$$B\left(\frac{\partial \Gamma}{\partial B}\right)^T = I_k$$

Proof: $\frac{\partial \Gamma}{\partial b_{ij}} = 1/2 tr\{(B \Lambda B^T)^{-1} \frac{\partial}{\partial b_{ij}} (B \Lambda B^T)\}$

$$= tr\{\frac{\partial B}{\partial b_{ij}} [\Lambda B^T (B \Lambda B^T)^{-1}]\}$$

so that the result follows.

Now, let Q be any nonsingular k by k matrix; the following Lemma follows immediately from Lemmas 1 and 2.

Lemma 3 If $\hat{B} = QB$, where Q is any nonsingular k by k matrix, then

$$\left(\frac{\partial \psi}{\partial \hat{B}}\right)^T = \left(\frac{\partial \psi}{\partial B}\right)^T Q^{-1}$$

and

$$\left(\frac{\partial \Gamma}{\partial \hat{B}}\right)^T = \left(\frac{\partial \Gamma}{\partial \hat{B}}\right)^T Q^{-1}$$

We now define the average divergence D_B and compute $\frac{\partial D_B}{\partial B}$. Assume the existence of m-distinct classes, normally distributed with means and covariances:

μ_i n-dimensional mean vector for class i.

Λ_i n by n covariance matrix for class i, assumed to be positive definite.

Let $\delta_{ij} = \mu_i - \mu_j$ so that $\delta_{ij} \delta_{ij}^T = \delta_{ji} \delta_{ji}^T$

The interclass divergence between classes i and j is defined [Kullback, 1968] as

$$D(i,j) = 1/2 \; tr\{\Lambda_i^{-1}(\Lambda_j + \delta_{ij} \delta_{ij}^T)\} + 1/2 tr\{\Lambda_j^{-1}(\Lambda_i + \delta_{ij} \delta_{ij}^T)\} - n$$

Note that when $\Lambda_i = \Lambda_j$ and $\mu_i = \mu_j$,

$$D(i,j) = 0$$

so that D(i,j) is in a sense, a measure of the degree of difficulty of distinguishing between classes i and j, with the larger the value of D(i,j), the less the degree of difficulty of distinguishing between classes i and j. We define the average divergence D to be

$$D = c \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} D(i,j)$$

$$= \frac{c}{2} \mathrm{tr}\{ \sum_{i=1}^{m} \Lambda_i^{-1} (\sum_{\substack{j=1 \\ j \neq i}}^{m} [\Lambda_j + \delta_{ij} \delta_{ij}^T]) \} - n$$

$$= \frac{c}{2} \mathrm{tr}\{ \sum_{i=1}^{m} \Lambda_i^{-1} S_i \} - n$$

where

$$S_i = \sum_{\substack{j=1 \\ j \neq i}}^{m} [\Lambda_j + \delta_{ij} \delta_{ij}^T]$$

$$c = 2/(m^2 - m)$$

Let

x ; an n-dimensional observation vector

B ; a k by n matrix of rank k, with $k \leq n$, and

$y = Bx$; the k-dimensional transformed observation vector.

The m transformed classes are also normally distributed and satisfy

$B\mu_i$; k-dimensional mean vector for class i

$B\Lambda_i B^T$; k by k covariance matrix for class i, which is positive definite by the assumptions on B and Λ_i.

Thus, in the range space of B, the B-induced interclass divergence $D_B(i.j)$, is, by definition of the interclass divergence;

$$D_B(i,j) = 1/2 \ \mathrm{tr}\{(B\Lambda_i B^T)^{-1} B(\Lambda_j + \delta_{ij} \delta_{ij}^T) B^T\}$$

$$+ 1/2 \ \mathrm{tr}\{(B\Lambda_j B^T)^{-1} B(\Lambda_i + \delta_{ij} \delta_{ij}^T) B^T\} - k$$

Similarly we can define the B-average divergence, D_B, as

$$D_B = c \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} D_B(i,j)$$

$$= \frac{c}{2} \mathrm{tr}\{ \sum_{i=1}^{m} [(B\Lambda_i B^T)^{-1} (BS_i B^T)] \} - k$$

where, as defined previously

$$S_i = \sum_{\substack{j=1 \\ j \neq i}}^{m} [\Lambda_j + \delta_{ij} \delta_{ij}^T]$$

is constant and need be computed but once if D_B is to be maximized numerically.

Note that in performing the transformation $y = Bx$, the dimension of each observation is reduced from n to k, so that in a sense, information is lost. A measure of the information lost is given by the difference

$$D - D_B \geq 0$$

which is necessarily non-negative (Kullback, 1968). We remark that if $D_B = D$, then B is said to be a sufficient statistic for the average divergence. We are interested in minimizing the information lost, as measured by the average divergence. Thus, it is desired to maximize the B-average divergence, or equivalently, minimize $-D_B$. We prove the following theorem:

Theorem 1 - (i) Each k by n matrix of rank k maximizing D_B must satisfy

$$\left(\frac{\partial D_B}{\partial B}\right)^T = c \sum_{i=1}^{m} [S_i B^T - \Lambda_i B^T (B \Lambda_i B^T)^{-1} (B S_i B^T)](B \Lambda_i B^T)^{-1} = (0)$$

(ii) If $\hat{B} = QB$, where Q is a nonsingular k by k matrix

$$\left(\frac{\partial D_{\hat{B}}}{\partial \hat{B}}\right)^T = \left(\frac{\partial D_B}{\partial B}\right)^T Q^{-1} \text{ and } D_{\hat{B}} = D_B$$

(iii) There exists a k by n matrix B of rank k that maximizes D_B.

Proof: The proof of (i) and (ii) follows from the definitions of B and D_B, and Lemmas 1 and 3. Since BB^T is positive definite, there exists a nonsingular k by k transformation Q such that $(QB)(QB)^T = I_k$, and since $D_{QB} = D_B$ by (ii) it suffices to consider only those matrices B belonging to the set $\beta = [B \mid BB^T = I_k]$. But β is a compact subset of E^{kn}. The proof of (iii) follows by noting that D_B is a continuous scaler valued function on the compact set β.

Thus, the problem of maximizing D_B amounts to determining kn distinct elements b_{ij} such that if $B = \{b_{ij}\}$, then $\frac{\partial D_B}{\partial B} = (0)$. We show that it is only necessary to determine k(n-k) distinct elements, for assume B is of the form

$$B = (I_k \vdots S)$$

where S is a k by n-k matrix so that we write $\left(\frac{\partial D_B}{\partial s_{ij}}\right) = \frac{\partial D_B}{\partial S}$ to represent the partial derivative of the B-average divergence with respect to the matrix S, evaluated at $B = (I_k \vdots S)$.

Corollary 1.1 - If $B = (I_k \vdots S)$, then $\frac{\partial D_B}{\partial S} = (0)$ implies $\frac{\partial D_B}{\partial B} = (0)$

Proof: Immediate, since for all B, by Theorem 1, $B\left(\frac{\partial D_B}{\partial B}\right)^T = (0)$

Lemmas 1-3 enable us to briefly investigate the Bhattacharyya Distance (Kailath, 1967) for two multivariate normal distributions. The Bhattacharyya Distance between two classes $\pi_1 = \pi_1(\mu_1, \Lambda_1)$ and $\pi_2 = \pi_2(\mu_2, \Lambda_2)$ is defined as

$$R(1,2) = 1/8 \; tr\{\left[\frac{\Lambda_1 + \Lambda_2}{2}\right]^{-1} \delta_{12} \delta_{12}^T\} + H(1,2)$$

where

$$H(1,2) = 1/2 \log \left|\frac{\Lambda_1 + \Lambda_2}{2}\right| - 1/4 \log |\Lambda_1 \Lambda_2|$$

The Bhattacharyya Distance is perhaps desirable as a feature selection criterion, since a bound on the probability of misclassification can readily be obtained from R(1,2)(Kailath, 1967). It follows that the transformed Bhattacharyya Distance resulting from the transformation $y = Bx$ is given by

$$R_B(1,2) = 1/8 \text{ tr}\left\{\left[\frac{B(\Lambda_1 + \Lambda_2)B^T}{2}\right]^{-1} B(\delta_{12}\delta_{12}^T)B^T\right\} + H_B(1,2)$$

where

$$H_B(1,2) = 1/2 \log \left|\frac{B(\Lambda_1 + \Lambda_2)B^T}{2}\right| - 1/4 \log \left|(B\Lambda_1 B^T)(B\Lambda_2 B^T)\right|$$

We prove

Theorem 2 - Let a k by n matrix B of rank k extremize the transformed Bhattacharyya Distance $R_B(1,2)$. Then it is necessary B satisfy an equation of the form

$$\left(\frac{\partial R_B(1,2)}{\partial B}\right)^T = \tfrac{1}{2}[\delta_{12}\delta_{12}^T B^T - (\Lambda_1+\Lambda_2)B^T[B(\Lambda_1+\Lambda_2)B^T]^{-1}(B\delta_{12}\delta_{12}^T B^T)\}[B(\Lambda_1+\Lambda_2)B^T]^{-1}$$

$$+ (\Lambda_1+\Lambda_2)B^T[B(\Lambda_1+\Lambda_2)B^T]^{-1} - \tfrac{1}{2}[\Lambda_1 B^T(B\Lambda_1 B^T)^{-1} + \Lambda_2 B^T(B\Lambda_2 B^T)^{-1}]$$

$$= (0)$$

Also, if $\hat{B} = QB$, where Q is a nonsingular k by k matrix, then

$$\left(\frac{\partial R_{\hat{B}}(1,2)}{\partial \hat{B}}\right)^T = \left(\frac{\partial R_B(1,2)}{\partial B}\right)^T Q^{-1}$$

Proof: Immediate by Lemmas 1, 2, and 3.

Corollary 2.1 - If $B = (I_k \vdots S)$, then $\frac{\partial R_B(1,2)}{\partial S} = (0)$ implies $\frac{\partial R_B(1,2)}{\partial B} = (0)$

Proof: Immediate, since for all B, $B\left(\frac{\partial R_B(1,2)}{\partial B}\right)^T = 0$ by Theorem 2.

Theorems 1 and 2 reveal that in maximizing either D_B or $R_B(1,2)$, it suffices to consider only those B satisfying $BB^T = I_k$ (i.e., for any rank k matrix B, there always exists a nonsingular k by k matrix Q satisfying $(QB)(QB)^T = I_k$). Since $\beta = \{B \mid BB^T = I_k\}$ is a compact subset of the class of k by n matrices (with the Euclidian topology), it follows that the maximum of D_B or $R_B(1,2)$ must necessarily be obtained on β. Moreover, given any $B \in \beta$, it is possible to construct an (n-k) by n matrix S satisfying $SS^T = I_{n-k}$, $BS^T = (0)$, and such that the n by n matrix

$$P = \begin{pmatrix} B \\ S \end{pmatrix}$$

satisfies $PP^T = I_n$, i.e., P is an orthogonal matrix. It follows that for any $B \in \beta$,

$$B = (I_k \vdots 0)\ P$$

and the solution to the feature selection problem amounts to optimally "rotating" or "reflecting" the original coordinates of the spectral measurement space (i.e., $X \to PX$) and then selecting the first k components of the resulting vector.

The following theorem [Bellman, 1970] is essential to the discussion and is included since it can be used to show $0 \le D_B \le D$ and $R_B(1,2) \le R(1,2)$ using only expressions previously defined in this section (i.e., by only using matrix algebra).

Theorem 3 - Consider the sequence of symmetric matrices

$$A_r = (a_{ij}) \qquad i,j=1,\ldots,r$$

for r=1,2,...,n. Let $\lambda_k(A_r)$ denote the k'th characteristic root of A_r, where

$$\lambda_1(A_r) \geq \lambda_2(A_r) \geq \ldots\ldots \geq \lambda_r(A_r)$$

Then

$$\lambda_{k+1}(A_{i+1}) \leq \lambda_k(A_i) \leq \lambda_k(A_{i+1})$$

Corollary 3.1 $\quad \lambda_{k+(n-i)}(A_n) \leq \lambda_k(A_i) \leq \lambda_k(A_n)$

We can use Corollary 3.1 to relate $tr\{B\Lambda B^T\}$ to $tr\{\Lambda\}$, where Λ is any n by n symmetric matrix and $BB^T = I_k$. Simply "extend" B to a nonsingular n by n matrix $P = \begin{pmatrix} B \\ S \end{pmatrix}$ where $P\,P^T = I_n$, the n by n identity matrix. The eigenvalues of $P\,\Lambda\,P^T$ are the same as those of Λ, so that Theorem 3 and Corollary 3.1 apply to $B\,\Lambda\,B^T$, considered as a submatrix of $P\,\Lambda\,P^T$. Thus, it follows that when $BB^T = I_k$, the trace of $B\,\Lambda\,B^T$ is bounded below by the sum of the k-smallest eigenvalues of Λ and is bounded above by the sum of the k-largest eigenvalues of Λ. Using this result, the following Theorems can be proved and are included for completeness.

Theorem 4 - Denote the n eigenvalues of

$$\Lambda_i^{-1} S_i \text{ by } \lambda_{i,1} \geq \lambda_{i,2} \geq \ldots\ldots \geq \lambda_{i,n}. \text{ Then}$$

$$\frac{c}{2} \sum_{i=1}^{m} \sum_{j=1}^{k} \lambda_{i,j+n-k} - k \leq D_B \leq \frac{c}{2} \sum_{i=1}^{m} \sum_{j=1}^{k} \lambda_{i,j} - k$$

Theorem 5 - Let $\lambda_1^2 \geq \lambda_2^2 \ldots\ldots \geq \lambda_\ell^2 \geq 1 > \lambda_{\ell+1}^2 \ldots\ldots \geq \lambda_n^2 > 0$ be the eigenvalues of $\Lambda_1^{-1}\,\Lambda_2$, and suppose that

$$\Phi_{max} = \{ \prod_{i=1}^{j} (\lambda_i + 1/\lambda_i) \} \{ \prod_{i=0}^{k-j-1} (\lambda_{n-i} + 1/\lambda_{n-i}) \}$$

maximizes the product of any k factors of the form $(\lambda_i + 1/\lambda_i)$; then

$$R_B(1,2) \leq 1/8\ \delta_{12}^T \left(\frac{\Lambda_1 + \Lambda_2}{2} \right)^{-1} \delta_{12} + 1/2 \log \frac{\Phi_{max}}{2} < R(1,2)$$

We remark again that Theorem 3.0 can also be used to show $D_B \leq D$ and $R_B(1,2) \geq 0$. We conclude this section by outlining a proof of the following Theorem which relates the average divergence D_B to the probability of misclassification PMC_B.

Theorem 6 - $D = D_B$ implies $PMC = PMC_B$

Proof: Let $p_i(x)$ be the density function for the i'th class, defined by mean μ_i and covariance Λ_i. Let $g_i(Bx)$ be the corresponding transformed density function for the i'th class defined by mean $B\mu_i$ and covariance $B\,\Lambda_i\,B^T$. The condition $D = D_B$ [Kullback, 1968] implies

$$\frac{p_i(x)}{p_j(x)} = \frac{g_i(Bx)}{g_j(Bx)} \quad i,j=1,\ldots,m \text{ almost everywhere}$$

so that a vector x is assigned to class i using a maximum likelihood classification procedure (with the $p_i(x)$) if and only if Bx is assigned to class i using a maximum likelihood classification (with the $g_i(Bx)$). The result follows by noting that for any measurable set R,

$$\int_R g_i(y)\,dy = \int_{B^{-1}(R)} p_i(x)\,dx$$

where $B^{-1}(R) = \{x \mid Bx \in R\}$.

3. NUMERICAL RESULTS

Based on the results of the last section, our feature selection criterion is stated simply as:

$$\max_{B} D_B$$

where B is a k by n matrix of rank k. Since $D - D_B \geq 0$, we will take the difference $D - D_B$ to be a measure of the information lost in performing the transformation $y = Bx$. The problem of maximizing D_B is handled numerically using the algorithm of Fletcher and Powell [1963], incorporated into a computer program essentially as documented by Johnson [1969]. The expression for the gradient of D_B is as in Theorem 1. Using the above, a sample problem is solved as discussed below.

The twelve dimensional (n=12) statistics for nine distinct classes, corresponding to nine distinct crops along the C1 Flight Line were obtained. For a fixed $k \leq n$, an initial B is obtained by exhaustively determining the B which maximizes D_B subject to the constraints $BB^T = I_k$ and $b_{ij} = b_{ij}$, where $B = \{b_{ij}\}$. Such a procedure constitutes what is called an "Exhaustive Search Procedure". Using the "best" B as obtained above for a first guess, D_B is maximized numerically with the results being presented by Figure 1 for various values of k. The bottom line corresponds to the first guess and the top line corresponds to the solution. The Figure indicates that essentially all the "information" is in a subspace of dimension 6 or less.

One can graphically display "separability" using what is called a "Class Separability to be Gained Map". Consider a coordinate system whose ordinate (for a given value of k) is D_B (i,j) where now B is assumed to maximize D_B. The abscissa is the value of D(i,j), in the original space and for a given i-j pair, represents the separability between classes i and j. Since $D(i,j) \geq D_B(i,j)$, the distance of a given point from the diagonal line $D(i,j) = D_B(i,j)$ represents the separability to be gained for that class pair. Thus, for a given class pair, its location along the abscissa is fixed, and as k increases, the point corresponding to that class pair can only move vertically toward the diagonal boundary. Obviously, for large enough k, all the points will lie on the diagonal boundary.

Figures 2 and 3 present Class Separability to be Gained Maps corresponding to k = 3 and k = 6, respectively. Each figure presents the class separation to be gained corresponding to the B obtained from the exhaustive search procedure (the initial B) and the B which maximizes D_B. In addition, it is possible to relate the results, i.e., $D - D_B$ to the probability of misclassification. This is essentially accomplished by analytically computing a bound on the probability of misclassification, computed in range space of the matrix B. The bound is obtained [Quirein, 1973] by considering a distinct linear discriminate function for each distinct class pair. For a fixed k, a bound is presented for the initial B obtained from the exhaustive search procedure and for that B which maximizes D_B obtained from the optimization program. From the figure, an "adequate" value of k is seen to be 3, so that for this particular problem, all the discriminatory information essentially lies in some three-dimensional subspace.

3. REFERENCES

Anderson, T. W., An Introduction to Multivariate Statistical Analysis, 1958 John Wiley and Sons, Inc., New York.

Babu, C. C., and Kalra, S. N., "On Feature Extraction in Multiclass Pattern Recognition," Int. J. Control, 1972, Vol. 15, No. 3.

Bellman, R., Introduction to Matrix Analysis, 1970 McGraw-Hill Book Company, New York.

Bond, A. C. "Feature Selection - The Without Replacement Procedure" TRW IOC 6534.6-72-72, 20 November 1972.

Fletcher, R., and Powell, J., "A Rapidly Converging Descent Method for Minimization," British Computer J., pp. 163-168, 1963

Johnson, Ivan "Impulsive Orbit Transfer Optimization by an Accelerated Gradient Method," Journal of Spacecraft and Rockets, Volume 6, No. 5 May 1969

Kailath, T., "The Divergence and Bhrattacharyya Distance Measures in Signal Selection," IEEE Transactions on Communication Theory, Vol. 15, No. 1, pp. 52-60, February 1967.

Kullback, Solomon, Information Theory and Statistics, 1968 Dover Publications, New York.

Quirein, J. A. "An Interactive Approach to the Feature Selection Classification Problem," TRW Systems Technical Note 99900-H019-R0-00, December 1972.

Quirein, J. A., "Admissible Linear Procedures and Thresholding," Mathematics Department, University of Houston Report #15, April 1973.

Tou, J. T., and Heydorn, R. P., 1967, in Computer and Information Sciences, Vol. 2, edited by J. T. Tou (New York: Academic Press)

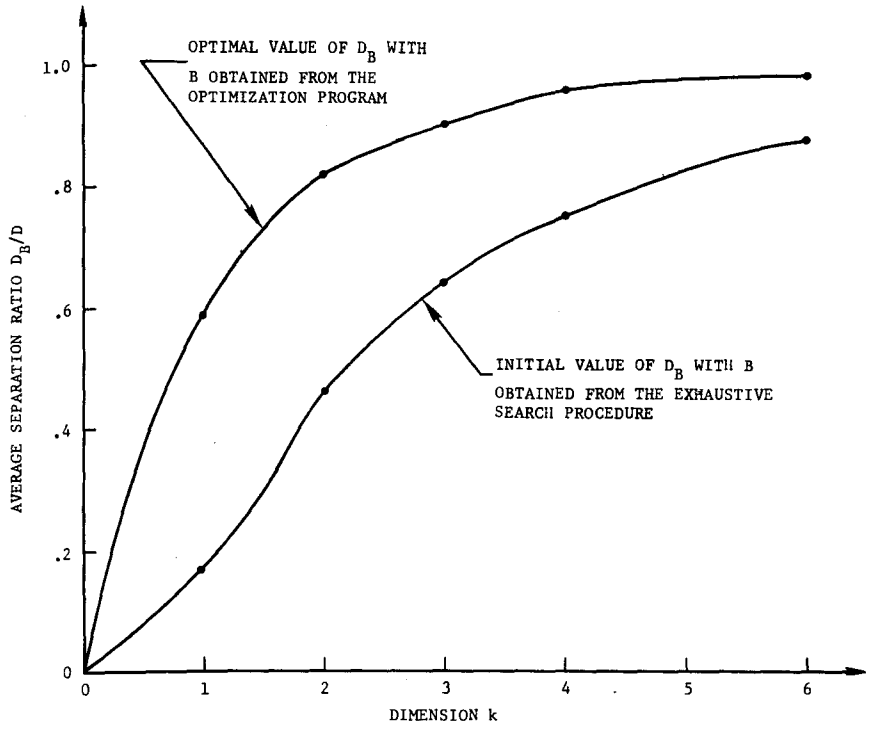

Figure 1. Numerical Results for Maximizing D_B.

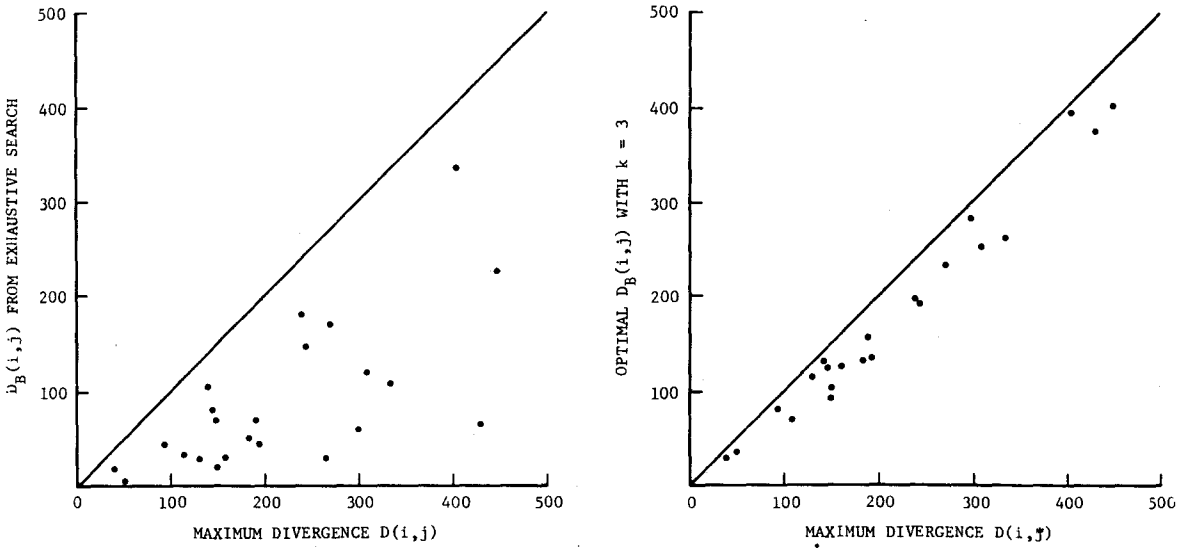

Figure 2. Class Separability to be Gained Maps (k=3).

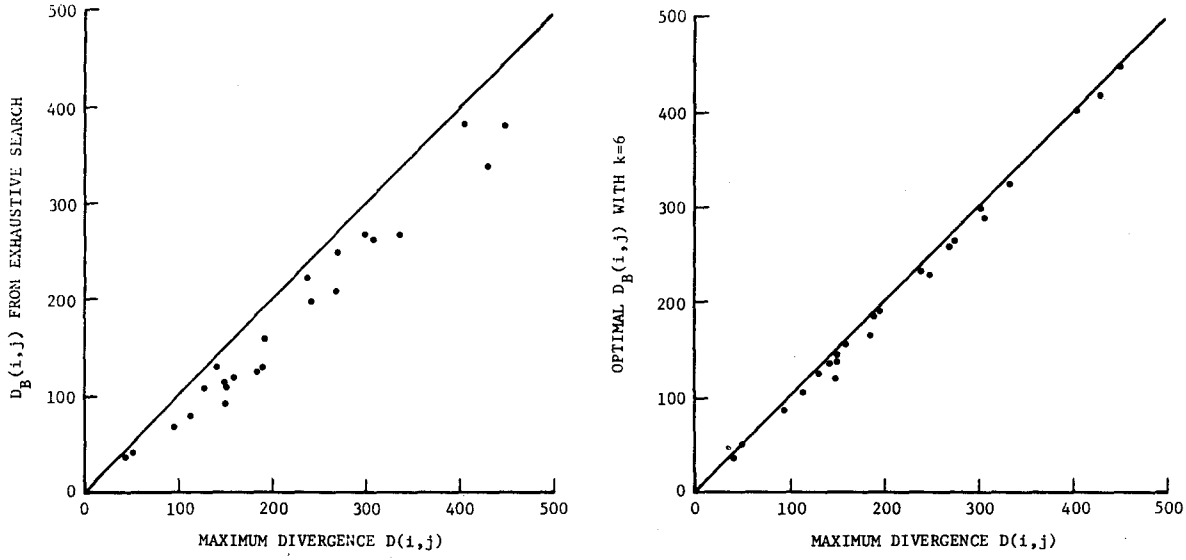

Figure 3. Class Separability to be Gained Maps (k=6).

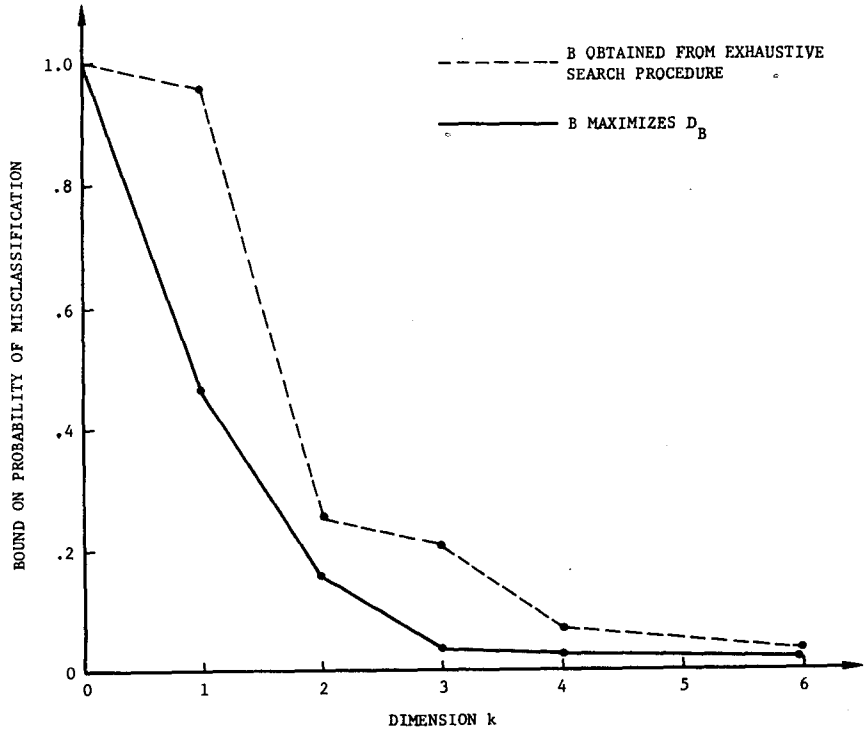

Figure 4. Bounding the Probability of Misclassification