Feature Extraction Based On Decision Boundaries

Chulhee Lee and David A. Landgrebe

School of Electrical Engineering
Purdue University, W. Lafayette, IN
Tel:(317)494-3846, FAX:(317)494-3358
landgreb@ecn.purdue.edu

FEATURE EXTRACTION BASED ON DECISION BOUNDARIES[1]

Chulhee Lee and David A. Landgrebe

School of Electrical Engineering
Purdue University, W. Lafayette, IN
Tel:(317)494-3846, FAX:(317)494-6440

< Abstract >

In this paper, a novel approach to feature extraction for classification is proposed based directly on the decision boundaries. We note that feature extraction is equivalent to retaining informative features or eliminating redundant features, thus first the terms "discriminantly informative feature" and "discriminantly redundant feature" are defined relative to feature extraction for classification. Next it is shown how discriminantly redundant features and discriminantly informative features are related to decision boundaries. A novel characteristic of the proposed method arises by noting that usually only a portion of the decision boundary is effective in discriminating between classes, and the concept of the effective decision boundary is therefore introduced. Next a procedure to extract discriminantly informative features based on a decision boundary is proposed. The proposed feature extraction algorithm has several desirable properties: (1) it predicts the minimum number of features necessary to achieve the same classification accuracy as in the original space for a given pattern recognition problem, and (2) it finds the necessary feature vectors. The proposed algorithm does not deteriorate under the circumstances of equal class means or equal class covariances as some previous algorithms do. Experiments show that the performance of the proposed algorithm compares favorably with those of previous algorithms.

[1] This work was funded in part by NASA under grant NAGW-925.

I. INTRODUCTION

Linear feature extraction can be viewed as finding a set of vectors that represent an observation while reducing the dimensionality. In pattern recognition, it is desirable to extract features that are focused on discriminating between classes. Although a reduction in dimensionality is desirable, the error increment due to the reduction in dimensionality must be constrained to be adequately small. Finding the minimum number of feature vectors which represent observations with reduced dimensionality without sacrificing the discriminating power of pattern classes along with finding the specific feature vectors has been one of the most important problems of the field of pattern analysis and has been studied extensively [1-12].

In this paper, we address this problem and propose a new algorithm for feature extraction based on the decision boundary. The algorithm predicts the minimum number of features to achieve the same classification accuracy as in the original space; at the same time the algorithm finds the needed feature vectors. Noting that feature extraction can be viewed as retaining informative features or eliminating redundant features, we define the terms discriminantly informative feature and discriminantly redundant feature. This reduces feature extraction to finding discriminantly informative features. We will show how discriminantly informative features and discriminantly redundant features are related to the decision boundary and can be derived from the decision boundary. We will need to define several terms and derive several theorems and, based on the theorems, propose a procedure to find discriminantly informative features from the decision boundary.

II. BACKGROUND AND PREVIOUS WORKS

Most linear feature extraction algorithms can be viewed as linear transformations. One of the most widely used transforms for signal representation is the Karhunen-Loeve transformation. Although the Karhunen-Loeve transformation is optimum for signal representation in the sense that it provides the smallest mean square error for a given number of features, quite often the features defined by the Karhunen-Loeve transformation are not optimum with regard to class separability.

In feature extraction for classification, it is not the mean square error but the classification accuracy that must be considered as a primary criterion for feature extraction.

Many authors have attempted to find the best features for classification based on criterion functions. Fisher's method finds the vector that gives the greatest class separation as defined by a criterion function [1]. Fisher's linear discriminant can be generalized to multiclass problems. In canonical analysis [2], a within-class scatter matrix Σ_W and a between-class scatter matrix Σ_b are used to formulate a criterion function and a vector $\mathbf{d}$ is selected to maximize,

$$\frac{\mathbf{d}^t\Sigma_b\mathbf{d}}{\mathbf{d}^t\Sigma_w\mathbf{d}}$$

where,

$$\Sigma_W = \sum_i P(\omega_i)\Sigma_i \qquad \text{(within-class scatter matrix)}$$

$$\Sigma_b = \sum_i P(\omega_i)(\mathbf{M}_i - \mathbf{M}_0)(\mathbf{M}_i - \mathbf{M}_0)^t \qquad \text{(between-class scatter matrix)}$$

$$\mathbf{M}_0 = \sum_i P(\omega_i)\mathbf{M}_i$$

Here $\mathbf{M}_i$, Σ_i , and $P(\omega_i)$ are the mean vector, the covariance matrix, and the prior probability of class ω_i, respectively. Although the vector found by canonical analysis performs well in most cases, there are several problems with canonical analysis. First of all, if there is little or no difference in mean vectors, feature vectors selected by canonical analysis is not reliable. Second, if a class has a mean vector very different from the mean vectors of the other classes, that class will be dominant in calculating the between-class scatter matrix, thus resulting in ineffective feature extraction.

Fukunaga recognized that the best representational features are not necessarily the best discriminating features and proposed a preliminary transformation [3]. The Fukunaga-Koontz method first finds a transformation matrix $\mathbf{T}$ such that,

$$\mathbf{T}[\mathbf{S}_1 + \mathbf{S}_2]\mathbf{T}^{-t} = \mathbf{I}$$

where $\mathbf{S}_i$ is the autocorrelation matrix of class ω_i.

Fukunaga showed that $\mathbf{TS}_1\mathbf{T}^{-t}$ and $\mathbf{TS}_2\mathbf{T}^{-t}$ have the same eigenvectors and all the eigenvalues are bounded by 0 and 1. It can be seen that the eigenvector with the largest differences in eigenvalues is the axis with the largest differences in variances. The Fukunaga-Koontz method will work well in problems where the covariance difference is dominant with little or no mean difference. However, by ignoring the information of mean difference, the Fukunaga-Koontz method is not suitable in the general case and could lead to irrelevant results [4].

Kazakos proposed a linear scalar feature extraction algorithm that minimizes the probability of error in discriminating between two multivariate normally distributed pattern classes [5]. By directly employing the probability of error, the feature extraction method finds the best single feature vector in the sense that it gives the smallest error. However, if more than one feature is necessary, it is difficult to generalize the method.

Heydorn proposed a feature extraction method by deleting redundant features where redundancy is defined in terms of a marginal distribution function [6]. The redundancy test uses a coefficient of redundancy. However, the method does not find a redundant feature vector unless the vector is in the direction of one of the original feature vectors even though the redundant feature vector could be detected by a linear transformation.

Feature selection using statistical distance measures has also been widely studied and successfully applied [7-8,15]. However, as the dimension of data increases, the combination of features to be examined increases exponentially, resulting in unacceptable computational cost. Several procedures to find a sub-optimum combination of features instead of the optimum combination of features have been proposed with a reasonable computational cost [8]. However, if the best feature vector or the best set of feature vectors is not in the direction of any original feature vector, more features may be needed to achieve the same performance.

Depending on the characteristics of the data, it has been shown that the previous feature extraction/selection methods can be applied successfully. However, it is also true that there are some cases in which the previous methods fail to find the best feature vectors or even good feature vectors, thus resulting in difficulty in choosing a suitable method to solve a particular problem. Although some authors addressed this problem [9-11], there is still another problem. One must determine, for a given problem, how many features must be selected to meet the requirement. More fundamentally, it is difficult with the previous feature extraction/selection algorithms to predict the intrinsic discriminant dimensionality, which is defined as the smallest number of features needed to achieve the same classification accuracy as in the original space for a given problem.

In this paper, we propose a different approach to the problem of feature extraction for classification. The proposed algorithm is based on decision boundaries directly. The proposed algorithm predicts the minimum number of features needed to achieve the same classification accuracy as in the original space for a given problem and finds the needed feature vectors, and it does not deteriorate when mean or covariance differences are small.

III. FEATURE EXTRACTION AND SUBSPACE

A. Feature Extraction and Subspace

Let **X** be an observation in the N-dimensional Euclidean space E^N. Then **X** can be represented by

$$\mathbf{X} = \sum_{i=1}^{N} a_i \alpha_i \qquad \text{where } \{\alpha_1, \alpha_2, .., \alpha_N\} \text{ is a basis of } E^N.$$

Then feature extraction is equivalent to finding a subspace, **W**, and the new features can be found by projecting an observation into the subspace. Let **W** be a M-dimensional subspace of E^N spanned by M linearly independent vectors, $\beta_1, \beta_2, .., \beta_M$.

$$\mathbf{W} = \text{Span}\{\beta_{1,}\beta_2, .., \beta_M\} \text{ and } \dim(\mathbf{W}) = M \leq N$$

Assuming that β_i's are orthonormal, the new feature set in subspace **W** is given by

$$\{\mathbf{X}^t\beta_1, \mathbf{X}^t\beta_2, .., \mathbf{X}^t\beta_M\} = \{b_1, b_2, .., b_M\} \text{ where } b_i = \mathbf{X}^t\beta_i$$

Now let $\hat{\mathbf{X}} = \sum_{i=1}^{M} b_i \beta_i$. Then $\hat{\mathbf{X}}$ will be an approximation to **X** in terms of a linear combination of $\{\beta_1, \beta_2, .., \beta_M\}$ in the original N-dimensional space.

B. Bayes' Decision Rule

Now consider briefly Bayes' decision rule, which will be used later in the proposed feature extraction algorithm. Let **X** be an observation in the N-dimensional Euclidean space E^N under hypothesis H_i: **X** ω_i i=1,2. Decisions will be made according to the following rule:

$$\text{Decide } \omega_1 \text{ if } P(\omega_1)P(\mathbf{X}|\omega_1) > P(\omega_2)P(\mathbf{X}|\omega_2) \qquad \text{else } \omega_2$$

where $P(\mathbf{X}|\omega_i)$ is a conditional density function.

Let $h(\mathbf{X}) = -\ln\frac{P(\mathbf{X}|\omega_1)}{P(\mathbf{X}|\omega_2)}$ and $t = \ln\frac{P(\omega_1)}{P(\omega_2)}$. Then

$$\text{Decide } \omega_1 \text{ if } h(\mathbf{X}) < t \qquad \text{else } \omega_2$$

Feature extraction has been used in many applications, and the criteria for feature extraction can be different in each case. If feature extraction is directed specifically at classification, a criterion could be to maintain classification accuracy. As a new approach to feature extraction for classification, we will find a subspace, **W**, with the minimum dimension M and the spanning vectors $\{\beta_i\}$ of the subspace such that for any observation $\mathbf{X}$

$$(h(\mathbf{X}) - t)(h(\hat{\mathbf{X}}) - t) > 0 \qquad (1)$$

where $\hat{\mathbf{X}}$ is an approximation of $\mathbf{X}$ in terms of a basis of subspace **W** in the original N-dimensional space. The physical meaning of (1) is that the classification result for $\hat{\mathbf{X}}$ is the same as the classification result of $\mathbf{X}$. In practice, feature vectors might be selected in such a way as to maximize the number of observations for which (1) holds with a constraint on the dimensionality of subspaces. In this paper, we will propose an algorithm which finds the minimum dimension of a subspace such that (1) holds for all the given observations and which also finds the spanning vectors $\{\beta_i\}$ of the subspace. In the next section, we define some needed terminology which will be used in deriving theorems later.

IV. DEFINITIONS

A. Discriminantly Redundant Feature

Feature extraction can be performed by eliminating redundant features, however, what is meant by "redundant" may be dependent on the application. For the purpose of feature extraction for classification, we will define a "discriminantly redundant feature" as follows.

Definition 1. Let $\{\beta_1, \beta_2, .., \beta_N\}$ be a orthonormal basis of E^N. We say the vector β_k is **discriminantly redundant** if for any observation $\mathbf{X}$

$$(h(\mathbf{X}) - t)(h(\hat{\mathbf{X}}) - t) > 0 \qquad (1)$$

In other words,

$$\text{if } h(\mathbf{X}) > t, \text{ then } h(\hat{\mathbf{X}}) > t \text{ or}$$

$$\text{if } h(\mathbf{X}) < t, \text{ then } h(\hat{\mathbf{X}}) < t$$

$$\text{where } \mathbf{X} = \sum_{i=1}^{N} b_i\beta_i,\ \hat{\mathbf{X}} = \sum_{i=1\ i\neq k}^{N} b_i\beta_i \text{ and } \mathbf{X} = \hat{\mathbf{X}} + b_k\beta_k$$

The physical meaning of (1) is that the classification result for $\hat{\mathbf{X}}$ is the same as the classification result of $\mathbf{X}$. Fig. 1 shows an example of a discriminantly redundant feature. In this case even though $\hat{\mathbf{X}}$ is moved along the direction of vector β_k, the classification result will remain unchanged. This means vector β_k makes no contribution in discriminating classes, thus vector β_k is redundant for the purpose of classification.

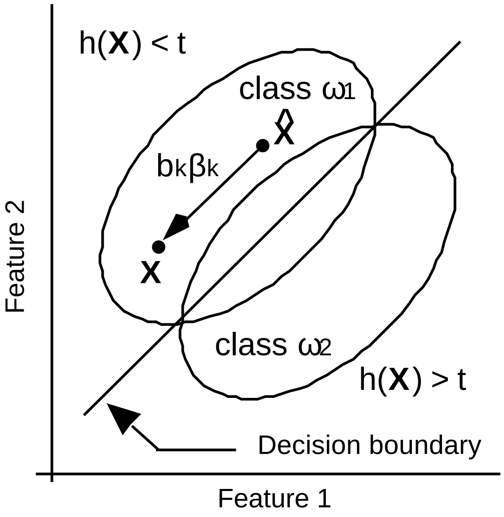

Fig. 1 An example of a discriminantly redundant feature. Even though the observation is moved in the direction of vector β_k, the decision will be the same.

B. Discriminantly Informative Feature

In a similar manner, we define a discriminantly informative feature.

Definition 2. Let $\{\beta_1,\beta_2,..,\beta_N\}$ be a orthonormal basis of E^N. We say that β_k is **discriminantly informative** if there exists an observation $\mathbf{Y}$ such that

$$(h(\mathbf{Y}) - t)(h(\hat{\mathbf{Y}}) - t) < 0 \qquad (2)$$

In other words,

$$h(\mathbf{Y}) > t \text{ but } h(\hat{\mathbf{Y}}) < t \text{ or}$$

$$h(\mathbf{Y}) < t \text{ but } h(\hat{\mathbf{Y}}) > t$$

$$\text{where } \mathbf{Y} = \sum_{i=1}^{N} b_i \beta_i,\ \hat{\mathbf{Y}} = \sum_{i=1\ i \neq k}^{N} b_i \beta_i \text{ and } \mathbf{Y} = \hat{\mathbf{Y}} + b_k \beta_k$$

The physical meaning of (2) is that there exists an observation **Y** such that the classification result of $\hat{\mathbf{Y}}$ is different from the classification result of **Y**. It is noted that (2) need not hold for all the observations. A vector will be discriminantly informative if there exists at least one observation whose classification result can be changed as the observation moves along the direction of the vector. Fig. 2 shows an example of a discriminantly informative feature. In this case, as $\hat{\mathbf{Y}}$ is moved along the direction of vector β_k the classification result will be changed.

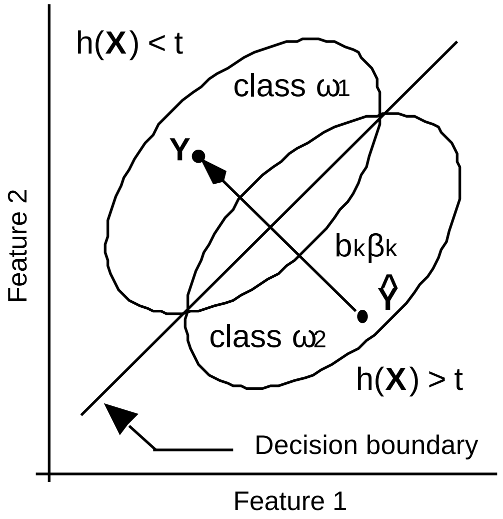

Fig. 2 An example of a discriminantly informative feature. As the observation **Y** is moved along the direction of vector $\boldsymbol{\beta}_k$, the classification result of the observation is changed.

C. Decision Boundaries and Effective Decision Boundaries

The decision boundary of a two-class problem is a locus of points on which *a posteriori* probabilities are the same. To be more precise, we define a decision boundary as follows:

Definition 3. A decision boundary is defined as

$$\{\ \mathbf{X} \mid h(\mathbf{X}) = t\ \}$$

A decision boundary can be a point, line, plane, hyper plane, solid, hyper solid, curved surface or curved hyper-surface. Although a decision boundary can be extended to infinity, in most cases some portion of the decision boundary is not significant. For practical purposes, we define the effective decision boundary as follows:

Definition 4. The effective decision boundary is defined as

$$\{ \mathbf{X} \mid h(\mathbf{X}) = t \,, \mathbf{X} \quad R_1 \text{ or } \mathbf{X} \quad R_2 \}$$

where R_1 is the smallest region which contains a certain portion, $P_{threshold}$, of class ω_1 and R_2 is the smallest region which contains a certain portion, $P_{threshold}$, of class ω_2.

The effective decision boundary may be seen as an intersection of the decision boundary and the regions where most of the data are located. Figs. 3 and 4 show some examples of decision boundaries and effective decision boundaries. In these examples, the threshold probability, $P_{threshold}$, is set to 99.9%. In the case of Fig. 3, the decision boundary is a straight line and the effective decision boundary is a straight line segment, the latter being a part of the former. In Fig. 4, the decision boundary is an ellipse and the effective decision boundary is a part of that ellipse which could be approximated by a straight line.

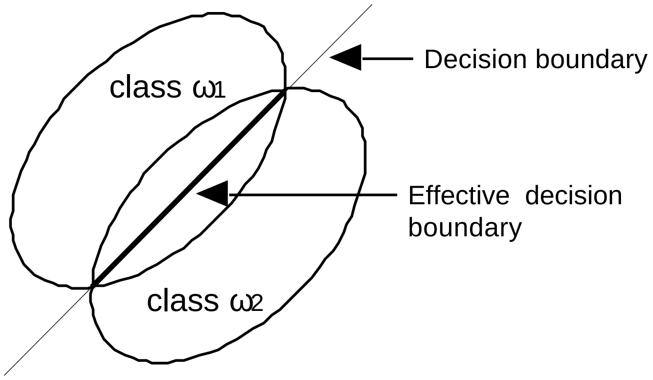

Fig. 3 $\mathbf{M}_1 \neq \mathbf{M}_2$, $\Sigma_1 = \Sigma_2$. The decision boundary is a straight line and the effective decision boundary is a line segment coincident to it.

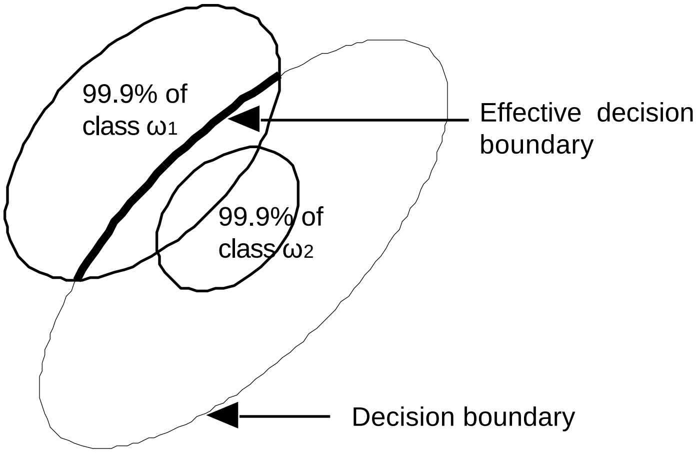

Fig. 4 $\mathbf{M}_1 \neq \mathbf{M}_2$, $\Sigma_1 \neq \Sigma_2$. The decision boundary is an ellipse and the effective decision boundary is a part of the ellipse which can be approximated by a straight line.

D. Intrinsic Discriminant Dimension

One of the major problems of feature extraction for classification is to find the minimum number of features needed to achieve the same classification accuracy as in the original space. To be more exact, we define the term, "intrinsic discriminant dimension".

Definition 5. The Intrinsic discriminant dimension for a given problem is defined as the smallest dimension of a subspace, $\mathbf{W}$, of the N-dimensional Euclidean space E^N such that for any observation $\mathbf{X}$ in the problem,

$$(h(\mathbf{X}) - t)(h(\hat{\mathbf{X}}) - t) > 0$$

$$\text{where } \mathbf{W} = \text{Span}\{\beta_1, \beta_2, .., \beta_M\},\ \hat{\mathbf{X}} = \sum_{i=1}^{M} b_i \beta_i \quad \mathbf{W} \text{ and } M \leq N.$$

The intrinsic discriminant dimension can be seen as the smallest dimensional subspace wherein the same classification accuracy can be obtained as could be obtained in the original space.

The intrinsic discriminant dimension is related to the discriminantly redundant feature vector and the discriminantly informative feature vector. In particular, if there are M linearly independent discriminantly informative feature vectors and L linearly independent discriminantly redundant feature vectors, then it can be easily seen that

$$N = M + L$$

where N is the original dimension and the intrinsic discriminant dimension is equal to M. Fig. 5 shows an example of the intrinsic discriminant dimension. In the case of Fig. 5, the intrinsic discriminant dimension is one even though the original dimensionality is two. If $\mathbf{V}_2$ is chosen as a new feature vector, the classification accuracy will be the same as in the original 2-dimensional space.

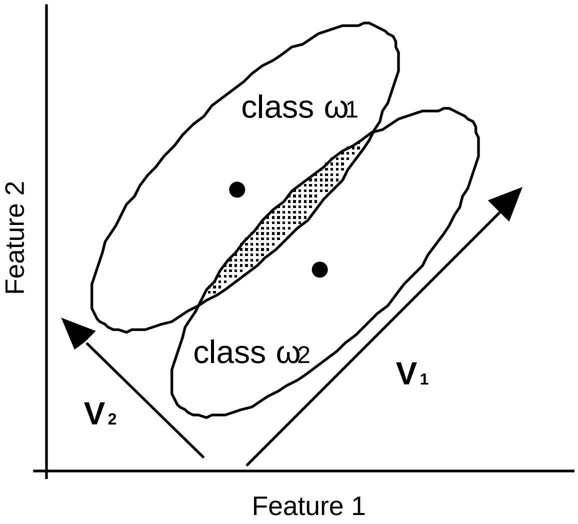

Fig. 5 $\Sigma_1=\Sigma_2$. In this case the intrinsic discriminant dimension is one even though the original space is two dimensional, since if $\mathbf{V}_2$ is chosen as a new feature vector, the classification accuracy will be the same as in the original 2 dimensional space.

V. FEATURE EXTRACTION BASED ON THE DECISION BOUNDARY

A. Redundancy Testing Theorem

From the definitions given in the previous section, a useful theorem can be stated which tests whether a feature vector is a discriminantly redundant feature or a discriminantly informative feature.

Theorem 1. If a vector is parallel to the tangent hyper-plane to the decision boundary at every point on the decision boundary for a pattern classification problem, then the vector contains no information useful in discriminating classes for the pattern classification problem, i.e., the vector is discriminantly redundant.

***Proof*.** Let $\{\beta_1,\beta_2,..,\beta_N\}$be a basis of the N-dimensional Euclidean space E^N, and let β_N be a vector that is parallel to the tangent hyper-plane to the decision boundary at every point on the decision boundary. Let **W** be a subspace spanned by N-1 spanning vectors, $\beta_1,\beta_2,..,\beta_{N-1}$, i.e.,

$$\mathbf{W} = \text{Span}\{\beta_1,\beta_2,..,\beta_{N-1}\} \text{ and } \dim(\mathbf{W}) = N-1$$

If β_N is not a discriminantly redundant feature vector, there must exist an observation $\mathbf{X}$ such that

$$(h(\mathbf{X}) - t)(h(\hat{\mathbf{X}}) - t) < 0$$

$$\text{where } \mathbf{X} = \sum_{i=1}^{N} b_i\beta_i \text{ and } \hat{\mathbf{X}} = \sum_{i=1}^{N-1} c_i\beta_i .$$

Without loss of generality, we can assume that the set of vectors $\beta_1,\beta_2,..,\beta_N$ is an orthonormal set. Then $b_i = c_i$ for i=1,N-1. Assume that there is an observation $\mathbf{X}$ such that

$$(h(\mathbf{X}) - t)(h(\hat{\mathbf{X}}) - t) < 0$$

This means $\mathbf{X}$ and $\hat{\mathbf{X}}$ are on different sides of the decision boundary. Then the vector

$$\mathbf{X}_d = \mathbf{X} - \hat{\mathbf{X}} = b_N\beta_N$$

where b_N is a coefficient, must pass through the decision boundary. But this contradicts the assumption that β_N is parallel to the tangent hyper-plane to the decision boundary at every point on the decision boundary. Therefore if β_N is a vector parallel to the tangent hyper-plane to the decision boundary at every point on the decision boundary, then for all observations $\mathbf{X}$

$$(h(\mathbf{X}) - t)(h(\hat{\mathbf{X}}) - t) > 0$$

Therefore β_N is discriminantly redundant. Fig. 6 shows an illustration of the proof.

From the theorem, we can easily derive the following lemmas which are very useful in finding discriminantly informative features.

Lemma 1. If vector $\mathbf{V}$ is orthogonal to the vector normal to the decision boundary at every point on the decision boundary, vector $\mathbf{V}$ contains no information useful in discriminating classes, i.e., vector $\mathbf{V}$ is discriminantly redundant.

Lemma 2. If a vector is normal to the decision boundary at at least one point on the decision boundary, the vector contains information useful in discriminating classes, i.e., the vector is discriminantly informative.

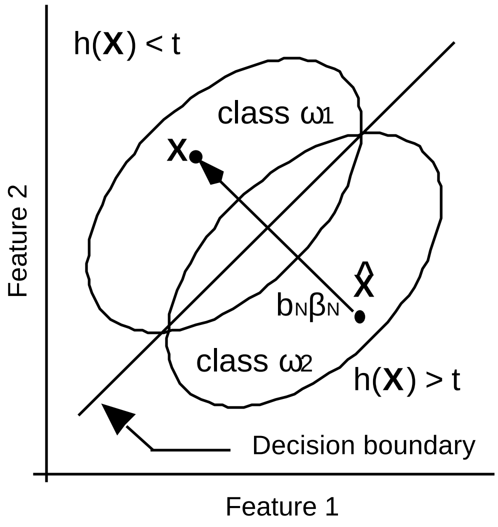

Fig. 6 If two observations are on the different sides of the decision boundary, the line connecting the two observations will pass through the decision boundary.

B. Decision Boundary Feature Matrix

From the previous theorem and lemmas, it can be seen that a vector normal to the decision boundary at a point is a discriminantly informative feature, and the effectiveness of the vector is roughly proportional to the area of the decision boundary which has the same normal vector. Now we can define a DECISION BOUNDARY FEATURE MATRIX which is very useful to predict the intrinsic discriminant dimension and find the necessary feature vectors.

Definition 6. The decision boundary feature matrix (DBFM): Let $\mathbf{N}(\mathbf{X})$ be the unit normal vector to the decision boundary at a point $\mathbf{X}$ on the decision boundary for a given pattern classification problem. Then the decision boundary feature matrix Σ_{DBFM} is defined as

$$\Sigma_{\text{DBFM}} = \frac{1}{K}\int_S \mathbf{N}(\mathbf{X})\mathbf{N}^t(\mathbf{X})p(\mathbf{X})d\mathbf{X}$$

where $p(\mathbf{X})$ is a probability density function, $K=\int_S p(\mathbf{X})d\mathbf{X}$, and S is the decision boundary, and the integral is performed over the decision boundary.

We will show some examples of the decision boundary feature matrices. Even though the examples are in 2-dimensional space, the concepts can be easily extended to higher dimensional spaces. In all examples, a Gaussian Maximum Likelihood classifier is assumed.

Example 1. The mean vectors and covariance matrices of two bivariate Gaussian classes are given as follows:

$$\mathbf{M}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \mathbf{S}_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \qquad \mathbf{M}_2 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad \mathbf{S}_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

$$P(\omega_1) = P(\omega_2) = 0.5$$

These distributions are shown in Fig. 7 as "ellipse of concentration." In a two-class, two-dimensional pattern classification problem, if the covariance matrices are the same, the decision boundary will be a straight line and the intrinsic discriminant dimension is one. This suggests that the vector normal to the decision boundary at any point is the same. And the decision boundary feature matrix will be given by

$$\Sigma_{DBFM} = \frac{1}{K} \int_S \mathbf{N}(\mathbf{X})\mathbf{N}^t(\mathbf{X})p(\mathbf{X})d\mathbf{X} = \frac{1}{K}\mathbf{N}\mathbf{N}^t \int_S p(\mathbf{X})d\mathbf{X} = \mathbf{N}\mathbf{N}^t$$

$$\Sigma_{DBFM} = \frac{1}{\sqrt{2}}(-1,1)^t \frac{1}{\sqrt{2}}(-1,1) = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

$$\text{Rank}(\Sigma_{DBFM}) = 1$$

It is noted that the rank of the decision boundary feature matrix is one which is equal to the intrinsic discriminant dimension and the eigenvector corresponding to the non-zero eigenvalue is the desired feature vector which gives the same classification accuracy as in the original 2-dimensional space.

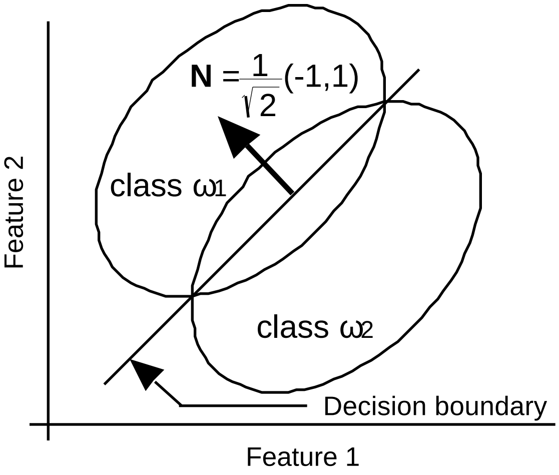

Fig. 7 An example where the covariance matrices of two classes are the same and the decision boundary is a straight line. In this case, there is only one vector which is normal to the decision boundary.

Example 2. The mean vectors and covariance matrices of two classes are given as follows:

$$\mathbf{M}_1 = \begin{bmatrix} 5 \\ 5 \end{bmatrix}, \quad \mathbf{S}_1 = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix} \qquad \mathbf{M}_2 = \begin{bmatrix} 5 \\ 5 \end{bmatrix}, \quad \mathbf{S}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$P(\omega_1) = P(\omega_2) = 0.5$$

The distributions of the two classes are shown in Fig. 8 as "ellipse of concentration." In this example, the decision boundary is a circle and symmetric, and $\frac{1}{K}p(\mathbf{X})$ is a constant given by $\frac{1}{2\pi r}$ where r is the radius of the circle. The decision boundary feature matrix will be given by

$$\Sigma_{\text{DBFM}} = \int_0^{2\pi} \frac{1}{2\pi r}[\cos\theta \ \sin\theta]^t[\cos\theta \ \sin\theta] \ r \ d\theta$$

$$= \frac{1}{2\pi}\int_0^{2\pi} \begin{bmatrix} \cos\theta\cos\theta & \cos\theta\sin\theta \\ \sin\theta\cos\theta & \sin\theta\sin\theta \end{bmatrix} d\theta$$

$$= \frac{1}{2\pi}\begin{bmatrix} \pi & 0 \\ 0 & \pi \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\text{Rank}(\Sigma_{\text{DBFM}}) = 2$$

From the distribution of data, it is seen that two features are needed to achieve the same classification accuracy as in the original space. This means that the intrinsic discriminant dimension is 2 in this case. It is noted that the rank of the decision boundary feature matrix is also 2, which is equivalent to the intrinsic discriminant dimension.

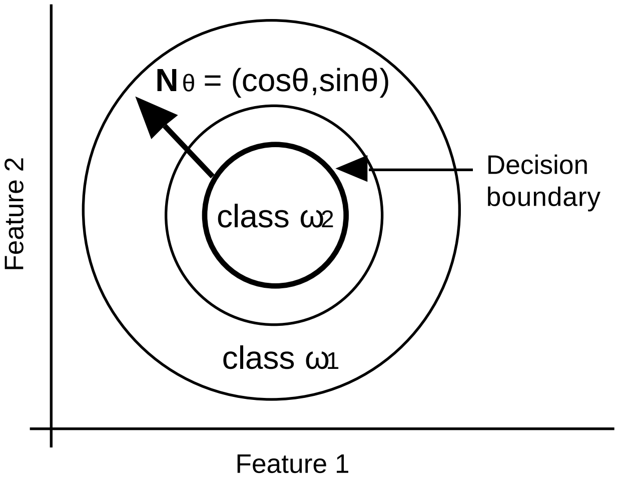

Fig. 8 The case for equal means and different covariances. In this case, the decision boundary will be a circle and there are an infinite number of different vectors which are normal to the decision boundary.

In a similar way, we define an EFFECTIVE DECISION BOUNDARY FEATURE MATRIX. The effective decision boundary feature matrix is the same as the decision boundary feature matrix except that only the effective decision boundary instead of the entire decision boundary is considered.

Definition 7. The effective decision boundary feature matrix (EDBFM): Let $\mathbf{N}(\mathbf{X})$ be the unit normal vector to the decision boundary at a point $\mathbf{X}$ on the effective decision boundary for a given pattern classification problem. Then the effective decision boundary feature matrix Σ_{EDBFM} is defined as

$$\Sigma_{\text{EDBFM}} = \frac{1}{K'} \int_{S'} \mathbf{N}(\mathbf{X})\mathbf{N}^t(\mathbf{X})p(\mathbf{X})d\mathbf{X}$$

where $p(\mathbf{X})$ is a probability density function, $K' = \int_{S'} p(\mathbf{X})d\mathbf{X}$, and S' is the effective decision boundary as defined in Definition 4, and the integral is performed over the effective decision boundary.

C. Decision Boundary Feature Matrix for Finding the Intrinsic Discriminant Dimension and Feature Vectors

We state the following two theorems which are useful in predicting the intrinsic discriminant dimension of a pattern classification problem and finding the feature vectors.

Theorem 2. The rank of the decision boundary feature matrix Σ_{DBFM} (Definition 6) of a pattern classification problem will be the intrinsic discriminant dimension (Definition 5) of the pattern classification problem.

Proof: Let **X** be an observation in the N-dimension Euclidean space E^N under the hypothesis H_i: **X** ω_i i = 1, 2. Let Σ_{DBFM} be the decision boundary feature matrix as defined in Definition 6. Suppose that

$$\text{rank}(\Sigma_{DBFM}) = M \le N.$$

Let $\{\varphi_1, \varphi_2,.., \varphi_M\}$ be the eigenvectors of Σ_{DBFM} corresponding to non-zero eigenvalues. Then a vector normal to the decision boundary at any point on decision boundary can be represented by a linear combination of φ_i, i=1,..,M. In other words, for any normal vector $\mathbf{V}_N$ to the decision boundary

$$\mathbf{V}_N = \sum_{i=1}^{M} a_i \varphi_i$$

Since any linearly independent set of vectors from a finite dimensional vector space can be extended to a basis for the vector space, we can expand $\{\varphi_1, \varphi_2,.., \varphi_M\}$ to form a basis for the N-dimension Euclidean space. Let $\{\varphi_1, \varphi_2,.., \varphi_M, \varphi_{M+1},.., \varphi_N\}$ be such a basis. Without loss of generality, we can assume $\{\varphi_1, \varphi_2,.., \varphi_M, \varphi_{M+1},.., \varphi_N\}$ is an orthonormal basis. One can always find an orthonormal basis for a vector space using the *Gram-Schmidt* procedure [13]. Since the basis is assumed to be orthonormal, it can be easily seen that the vectors $\{\varphi_{M+1}, \varphi_{M+2},.., \varphi_N\}$, are orthogonal to any vector $\mathbf{V}_N$ normal to the decision boundary. This is because for i = M+1,..,N

$$\varphi_i^t \mathbf{V}_N = \varphi_i^t \sum_{k=1}^{M} a_k \varphi_k$$

$$= \sum_{k=1}^{M} a_k \varphi_i^t \varphi_k = 0 \qquad \text{since } \varphi_i^t \varphi_k = 0 \text{ if } i \neq k$$

Therefore, since the vectors $\{\varphi_{M+1}, \varphi_{M+2},.., \varphi_N\}$ are orthogonal to any vector normal to the decision boundary, according to Lemma 1, the vectors $\{\varphi_{M+1}, \varphi_{M+2},.., \varphi_N\}$ are discriminantly redundant. Therefore the number of discriminantly redundant features is N – M, and the intrinsic discriminant dimension is M which is the rank of decision boundary feature matrix Σ_{DBFM}.

From Theorem 2 we can easily derive the following theorem which is useful to find the necessary feature vectors.

Theorem 3. The eigenvectors of the decision boundary feature matrix of a pattern recognition problem corresponding to non-zero eigenvalues are the necessary feature vectors to achieve the same classification accuracy as in the original space for the pattern recognition problem.

D. Procedure to Find the Decision Boundary Feature Matrix

Assuming a Gaussian Maximum Likelihood classifier is used, the decision boundary will be a quadratic surface if the covariance matrices are different. In this case, the rank of the decision boundary feature matrix will be the same as the dimension of the original space except for some special cases. However, in practice, only a small portion of the decision boundary is significant. Therefore if the decision boundary feature matrix is estimated using only the significant portion of the decision boundary or the efficient decision boundary, the rank of the decision boundary feature matrix, equivalently the number of features, can be reduced substantially while achieving about the same classification accuracy.

More specifically, the significance of any portion of the decision boundary is related to how much accuracy can be achieved by utilizing that portion of the decision boundary. Consider the case of Fig. 9 which shows the two regions which contain 99.9% of each Gaussianly distributed class, along with the decision boundary and the effective decision boundary of 99.9%. Although in this example the threshold probability, $P_{threshold}$, is set to 99.9% arbitrarily, it can be set to any value depending on the application (See Definition 4). If only the effective decision boundary, which is displayed in bold, is retained, it is still possible to classify 99.9% of data from class ω_1

the same as if the whole decision boundary had been used, since the effective decision boundary together with the boundary of the region which contains 99.9% of class ω_1 can divide the data of class ω_1 into two groups in the same manner as if the whole decision boundary is used; less than 0.1% of data from class ω_1 may be classified differently.

Therefore, for the case of Fig. 9, the effective decision boundary displayed as a bold line plays a significant role in discriminating between the classes, while the part of the decision boundary displayed as a non-bold line does not contribute much in discriminating between the classes. Other portions of the decision boundary displayed as a dotted line would be very rarely used.

It is noted, however, that even though only the effective decision boundary is used for feature extraction, this does not mean that the portion outside of the effective regions does not have a decision boundary. The actual decision boundary is approximated by the extension of the effective decision boundary as shown in Fig. 9. As shall be seen, feature extraction based on the effective decision boundary instead of the complete decision boundary will result in fewer features while achieving nearly the same classification accuracy.

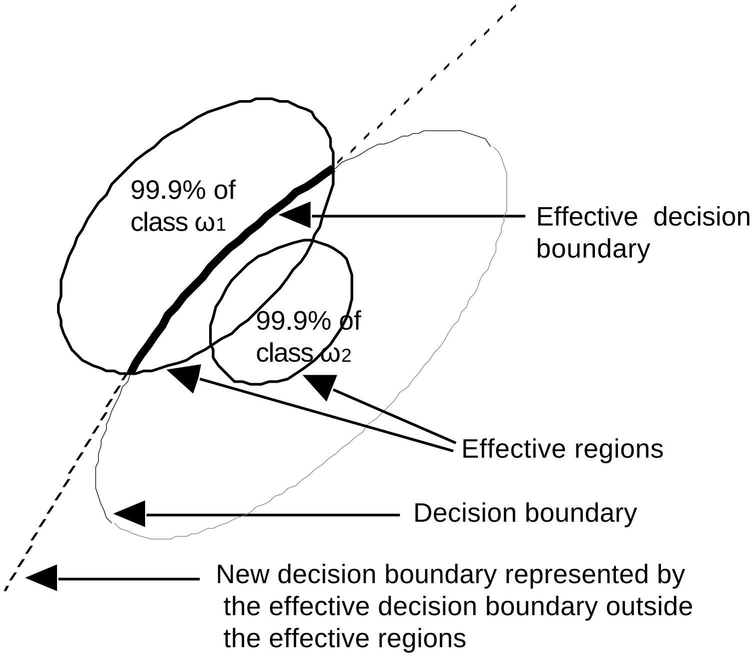
99.9% of
class ω1
99.9% of
class ω2
Effective decision boundary
Effective regions
Decision boundary
New decision boundary represented by the effective decision boundary outside the effective regions

Fig. 9 An example of a decision boundary and an effective decision boundary.

Next we propose a procedure for calculating the effective decision boundary feature matrix numerically.

Numerical Procedure to Find the Effective Decision Boundary Feature Matrix (2 pattern classes)

1. Let $\hat{\mathbf{M}}_i$ and $\hat{\Sigma}_i$ the estimated mean and covariance of class ω_i. Classify the training samples using full dimensionality. And apply a chi-square threshold test to the correctly classified training samples of each class and delete outliers. In other words, for class ω_i, retain **X** only if

$$(\mathbf{X} - \hat{\mathbf{M}}_i)^t \hat{\Sigma}_i^{-1} (\mathbf{X} - \hat{\mathbf{M}}_i) < R_{t1}$$

In the following STEPs, only correctly classified training samples which passed the chi-square threshold test will be used. Let $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{L_1}\}$ be such training samples of class ω_1 and $\{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{L_2}\}$ be such training samples of class ω_2.

For class ω_1, do STEP 2 through STEP 6:

2. Apply a chi-square threshold test of class ω_1 to the samples of class ω_2 and retain $\mathbf{Y}_j$ only if

$$(\mathbf{Y}_j - \hat{\mathbf{M}}_1)^t \hat{\Sigma}_1^{-1} (\mathbf{Y}_j - \hat{\mathbf{M}}_1) < R_{t2}$$

If the number of the samples of class ω_2 which pass the chi-square threshold test is less than L_{min} (see below), retain the L_{min} samples of class ω_2 which gives the smallest values.

3. For $\mathbf{X}_i$ of class ω_1, find the nearest sample of class ω_2 retained in STEP 2.
4. Find the point $\mathbf{P}_i$ where the straight line connecting the pair of samples found in STEP 3 meets the decision boundary.
5. Find the unit normal vector, $\mathbf{N}_i$, to the decision boundary at the point $\mathbf{P}_i$ found in STEP 4.
6. By repeating STEP 3 through STEP 5 for X_i, $i=1,..,L_1$, L_1 unit normal vectors will be calculated. From the normal vectors, calculate an estimate of the effective decision boundary feature matrix (Σ^1_{EDBFM}) from class ω_1 as follows:

$$\Sigma^1_{EDBFM} = \sum_i^{L_1} \mathbf{N}_i \mathbf{N}_i^t$$

Repeat STEP 2 through STEP 6 for class ω_2.

7. Calculate an estimate of the final effective decision boundary feature matrix as follows:

$$\Sigma_{EDBFM} = \Sigma^1_{EDBFM} + \Sigma^2_{EDBFM}$$

The chi-square threshold test in STEP 1 is necessary to eliminate outliers. Otherwise, outliers may give a false decision boundary when classes are well separable. The chi-square threshold test to the other class in STEP 2 is necessary to concentrate on effective decision boundary (Definition 4). Otherwise, the decision boundary feature matrix may be calculated from an insignificant portion of decision boundary, resulting in ineffective features. In the experiments, L_{min} in STEP 2 is set to 5 and R_{t1} is chosen such that

$$\Pr\{\mathbf{X} | (\mathbf{X} - \hat{\mathbf{M}}_i)^t \hat{\Sigma}_i^{-1} (\mathbf{X} - \hat{\mathbf{M}}_i) < R_{t1}\} = 0.95, \text{ i=1,2, and } R_{t1} = R_{t2}$$

Fig. 10 shows an illustration of the proposed procedure.

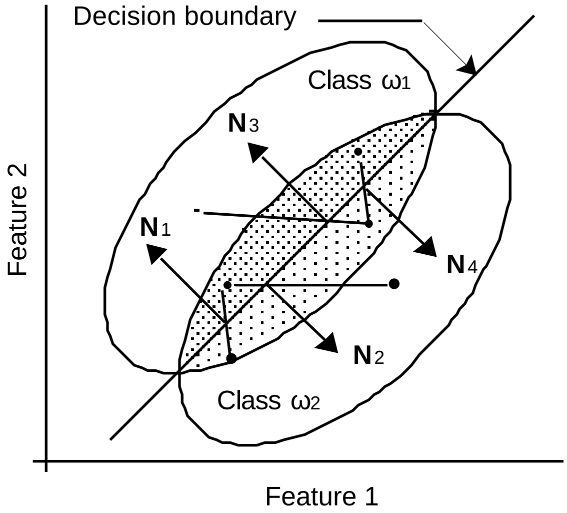

Fig. 10 Illustration of the procedure to find the effective decision boundary feature matrix numerically.

If we assume a Gaussian distribution for each class and the Gaussian ML classifier is used, $h(\mathbf{X})$ in Eq. (1) is given by

$$h(\mathbf{X}) = -\ln\frac{P(\mathbf{X}|\omega_1)}{P(\mathbf{X}|\omega_2)} = \frac{1}{2}(\mathbf{X} - \mathbf{M}_1)^t\Sigma_1^{-1}(\mathbf{X} - \mathbf{M}_1) + \frac{1}{2}\ln|\Sigma_1| - \frac{1}{2}(\mathbf{X} - \mathbf{M}_2)^t\Sigma_2^{-1}(\mathbf{X} - \mathbf{M}_2) - \frac{1}{2}\ln|\Sigma_2|$$

The vector normal to the decision boundary at $\mathbf{X}_0$ is given by [17]

$$\mathbf{N} = \quad h(\mathbf{X})|_{\mathbf{X}=\mathbf{X}_0} = (\Sigma_1^{-1} - \Sigma_2^{-1})\mathbf{X}_0 + (\Sigma_2^{-1}\mathbf{M}_1 - \Sigma_1^{-1}\mathbf{M}_2) \qquad (3)$$

If $\mathbf{P}_1$ and $\mathbf{P}_2$ are on different sides of decision boundary $h(\mathbf{X}) = t$, the point $\mathbf{X}_0$ where the line connecting $\mathbf{P}_1$ and $\mathbf{P}_2$ passes through the decision boundary is given by [17]

$$\mathbf{X}_0 = u\mathbf{V} + \mathbf{V}_0 \qquad (4)$$

where $\mathbf{V}_0 = \mathbf{P}_1$

$\mathbf{V} = \mathbf{P}_2 - \mathbf{P}_1$

$u = \dfrac{t - c'}{b}$ if a = 0,

$u = \dfrac{-b \pm \sqrt{b^2 - 4a(c' - t)}}{2a}$ and $0 \le u \le 1$ if $a \ne 0$,

$a = \dfrac{1}{2}\mathbf{V}^t(\Sigma_1^{-1} - \Sigma_2^{-1})\mathbf{V}$,

$b = \mathbf{V}_0{}^t(\Sigma_1^{-1} - \Sigma_2^{-1})\mathbf{V} - (\mathbf{M}_1^t\Sigma_1^{-1} - \mathbf{M}_2^t\Sigma_2^{-1})\mathbf{V}$,

$c' = \dfrac{1}{2}\mathbf{V}_0{}^t(\Sigma_1^{-1} - \Sigma_2^{-1})\mathbf{V}_0 - (\mathbf{M}_1^t\Sigma_1^{-1} - \mathbf{M}_2^t\Sigma_2^{-1})\mathbf{V}_0 + c$,

$c = \dfrac{1}{2}(\mathbf{M}^t\Sigma_1^{-1}\mathbf{M}_1 - \mathbf{M}_2^t\Sigma_2^{-1}\mathbf{M}_2) + \dfrac{1}{2}\ln\dfrac{|\Sigma_1|}{|\Sigma_2|}$

Eq. (4) can be used to calculate the point on the decision boundary from two samples classified differently and Eq. (3) can be used to calculate a normal vector to the decision boundary.

E. Decision Boundary Feature Matrix for Multiclass Problem

If there are more than two classes, the total decision boundary feature matrix can be defined as the sum of the decision boundary feature matrix of each pair of classes. If prior probabilities are available, the summation can be weighted. In other words, if there are M classes, the total decision boundary feature matrix can be defined as

$$\Sigma_{DBFM} = \sum_{i}^{M}\sum_{j,\, j\ne i}^{M} P(\omega_i)P(\omega_j)\Sigma_{DBFM}^{ij}$$

where Σ_{DBFM}^{ij} is the decision boundary feature matrix between class ω_i and class ω_j and $P(\omega_i)$ is the prior probability of class ω_i if available. Otherwise let $P(\omega_i)=1/M$.

It is noted that Theorem 2 and Theorem 3 still hold for multiclass case and the eigenvectors of the total decision boundary feature matrix corresponding to non-zero eigenvalues are the necessary feature vectors to achieve the same classification accuracy as in the original space. In practice, the total effective decision boundary feature matrix can be calculated by repeating the procedure for each pair of classes. By eliminating some redundancy, the total decision boundary feature matrix for a multiclass problem can be more efficient [17].

VI. EXPERIMENTS AND RESULTS

A. Experiments with synthetically generated data

To evaluate closely how the proposed algorithm performs under various circumstances, tests are conducted on data generated with given statistics assuming Gaussian distributions. In all examples, a Gaussian Maximum Likelihood classifier is used and the same data are used for training and test. In each example, the Foley & Sammon method [4] and the Fukunaga & Koontz method [3], are compared and discussed. In particular, classification accuracies of the Decision Boundary Feature Extraction method and the Foley & Sammon method are compared.

Example 3. In this example, data are generated for the following statistics.

$$\mathbf{M}_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \qquad \mathbf{M}_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

$$P(\omega_1) = P(\omega_2) = 0.5$$

Since the covariance matrices are the same, it can be easily seen that the decision boundary will be a straight line and just one feature is needed to achieve the same classification accuracy as in the original space. The eigenvalues λ_i and the eigenvectors φ_i of Σ_{EDBFM} are calculated as follows:

$$\lambda_1 = 576.97, \quad \lambda_2 = 0.03 \qquad \varphi_1 = \begin{bmatrix} 0.71 \\ -0.70 \end{bmatrix}, \quad \varphi_2 = \begin{bmatrix} 0.70 \\ 0.71 \end{bmatrix}$$

Since one eigenvalue is significantly larger than the other, it can be said that the rank of Σ_{EDBFM} is 1. That means only one feature is needed to achieve the same classification accuracy as in the original space. Considering the statistics of the two classes, the rank of Σ_{EDBFM} gives the correct number of features to achieve the same classification accuracy as in the original space. Fig. 11 shows the distribution of the generated data and the decision boundary found by the proposed procedure. Since class mean differences are dominant in this example, the Foley & Sammon method will also work well. However, the Fukunaga & Koontz method will fail to find the correct feature vector. With two features, the classification accuracy is 95.8% and both methods achieve the same accuracy (95.8%) with just one feature.

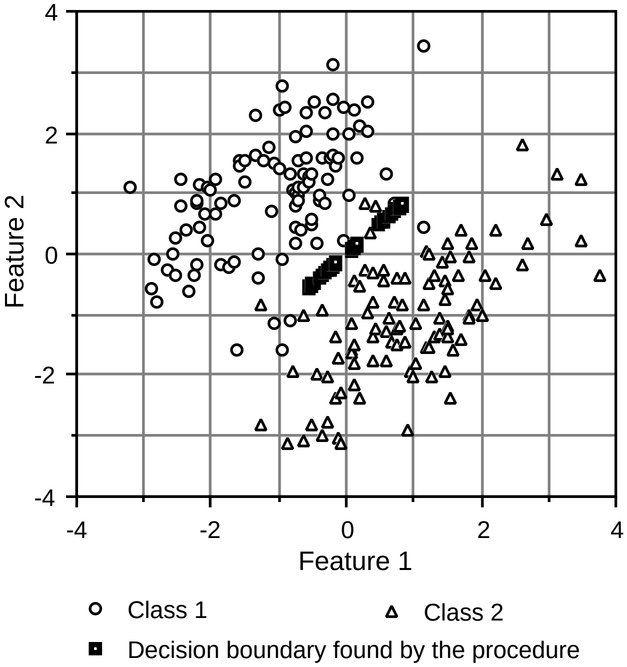

Fig. 11 The distribution of data for the two classes in Example 3. The decision boundary found by the proposed algorithm is also shown.

Example 4. In this example, data are generated with the following statistics.

$$\mathbf{M}_1 = \begin{bmatrix} 0.01 \\ 0 \end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix} \qquad \mathbf{M}_2 = \begin{bmatrix} -0.01 \\ 0 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$$

$$P(\omega_1) = P(\omega_2) = 0.5$$

In this case, there is almost no difference in the mean vectors and there is no correlation between the features for each class. The variance of feature 1 of class ω_1 is equal to that of class ω_2 while the variance of feature 2 of class ω_1 is larger than that of class ω_2. Thus the decision boundary will consist of hyperbolas, and two features are needed to achieve the same classification accuracy as in the original space. However, the effective decision boundary could be approximated by a straight line without introducing significant error. Fig. 12 shows the distribution of the generated data and the decision boundary obtained by the proposed procedure. The eigenvalues λ_i and the eigenvectors φ_i of Σ_{EDBFM} are calculated as follows:

$$\lambda_1 = 331.79, \quad \lambda_2 = 27.21 \qquad \varphi_1 = \begin{bmatrix} 0.06 \\ 1.00 \end{bmatrix}, \qquad \varphi_2 = \begin{bmatrix} -1.00 \\ 0.06 \end{bmatrix}$$

Since the rank of Σ_{EDBFM} is 2, two features are required to achieve the same classificationaccuracy as in the original space. However, λ_2 is considerably smaller than λ_1, even though λ_2 is not negligible. Therefore, nearly the same classification accuracy could be achieved with just one feature.

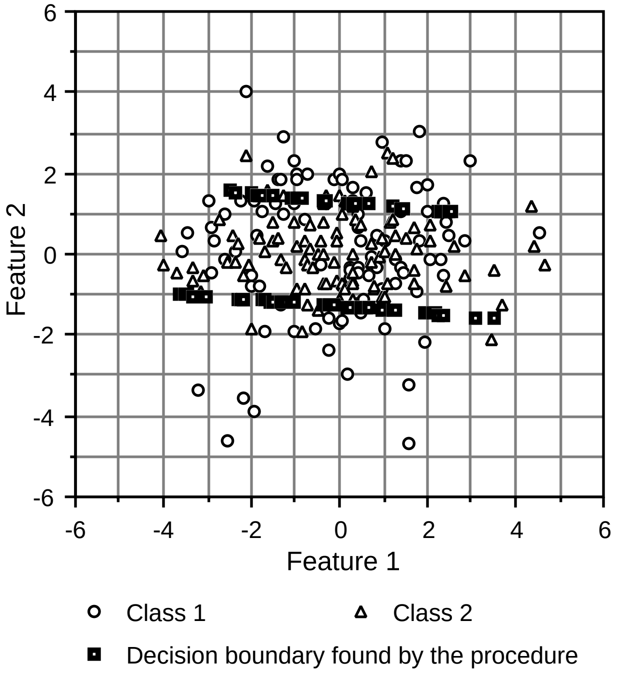

Fig. 12 Distribution of data from the two classes in Example 4. The decision boundary found by the proposed algorithm is also shown.

Since there is a very small difference in the mean vectors in this example, the Foley & Sammon method will fail to find the correct feature vector. On the other hand, the Fukunaga & Koontz method will find the correct feature vector. Table I shows classification accuracies. Decision Boundary Feature Extraction achieves the same accuracy with one feature as can be obtained with two features while the Foley & Sammon method fails to find the right feature in this example.

Table I. Classification accuracies of Decision Boundary Feature Extraction and the Foley & Sammon Method of Example 4.

No. Features	Decision Boundary Feature Extraction	Foley & Sammon Method
1	61.0 (%)	52.5 (%)
2	61.0 (%)	61.0 (%)

Example 5. In this example, we generate data for the following statistics.

$$\mathbf{M}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{M}_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$P(\omega_1) = P(\omega_1) = 0.5$$

In this case, there is no difference in the mean vectors and there are variance differences in only two features. It can be seen that the decision boundary will be a hollow right circular cylinder of infinite height and just two features are needed to achieve the same classification accuracy as in the original space. Eigenvalues λ_i and eigenvectors φ_i of Σ_{EDBFM} are calculated as follows:

$$\lambda_1 = 234.93\ ,\quad \lambda_2 = 171.49,\quad \lambda_3 = 1.58$$

$$\varphi_1 = \begin{bmatrix} 0.86 \\ -050 \\ 001 \end{bmatrix},\quad \varphi_2 = \begin{bmatrix} 0.49 \\ 0.84 \\ 0.21 \end{bmatrix},\quad \varphi_3 = \begin{bmatrix} -0.21 \\ -0.18 \\ 0.98 \end{bmatrix}$$

$$\mathrm{Rank}(\Sigma_{EDBFM}) \approx 2$$

Since the rank of Σ_{EDBFM} is 2, it can be said that two features are required to achieve the same classification accuracy as in the original space, which agrees with the data. Since there is no difference in the mean vectors in this example, the Foley & Sammon method will fail to find the correct feature vectors. On the other hand, the Fukunaga & Koontz method will find the correct feature vector. Table II shows the classification accuracies. Decision Boundary Feature Extraction finds the two effective feature vectors, achieving the same classification accuracy as in the original space.

Table II. Classification accuracies of Decision Boundary Feature Extraction and the Foley & Sammon Method of Example 5.

No. Features	Decision Boundary Feature Extraction	Foley & Sammon Method
1	65.0 (%)	62.3 (%)
2	70.0 (%)	60.5 (%)
3	70.0 (%)	70.0 (%)

B. Experiments with real data

In the following experiments, tests are conducted using multispectral data which was collected as a part of the LACIE remote sensing program [14] and major parameters are shown in Table III.

TABLE III. Parameters of Field Spectrometer System

Number of Bands	60 bands
Spectral Coverage	0.4 - 2.4 μm
Altitude	60 m
IFOV(ground)	25 m

Along with the proposed Decision Boundary Feature Extraction, three other feature selection/extraction algorithms CANONICAL ANALYSIS [2], feature selection using a statistical distance measure, and the Foley & Sammon method [4] are tested to evaluate and compare the performance of the proposed algorithm. In the feature selection using a statistical distance measure, Bhattacharyya distance [12] is used. Feature selection using the statistical distance measure will be referred as STATISTICAL SEPARABILITY. The Foley & Sammon method is based on the generalized Fisher criterion [4]. For a two class problem, the Foley & Sammon method is used for comparison. If there are more than 2 classes, CANONICAL ANALYSIS is used for comparison.

In the following test, two classes (WINTER WHEAT and UNKNOWN CROPS) are chosen from the data collected at Finney Co. KS. in May 3, 1977. WINTER WHEAT has 691 samples and UNKNOWN CROPS have 619 samples. In this test, the covariance matrices and mean vectors are estimated using 400 randomly chosen samples from each class and the rest of the data are used for test. Fig. 13 shows the mean graph of the two classes. There is reasonable difference in the mean vectors.

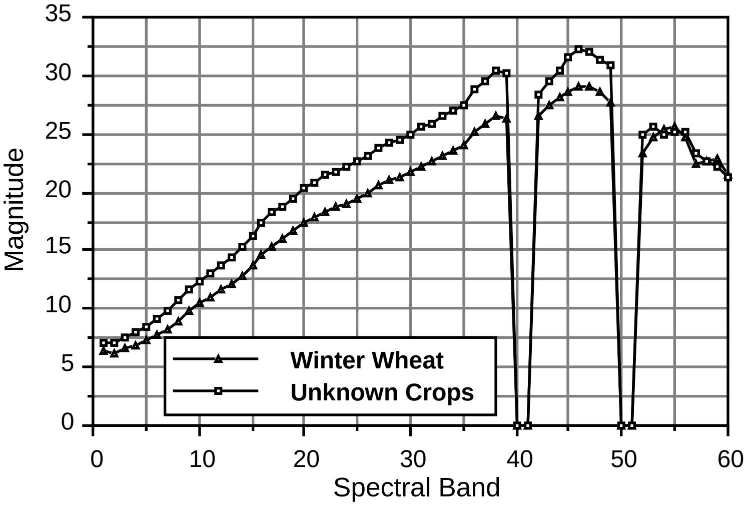

Fig. 13 Mean graph of the two classes.

Fig. 14 show the performance comparison of the 3 feature selection/extraction algorithms for different numbers of features. DECISION BOUNDARY FEATURE EXTRACTION and the Foley & Sammon method achieve approximately the maximum classification accuracy with just one feature while STATISTICAL SEPARABILITY needs 5 features to achieve about the same classification accuracy.

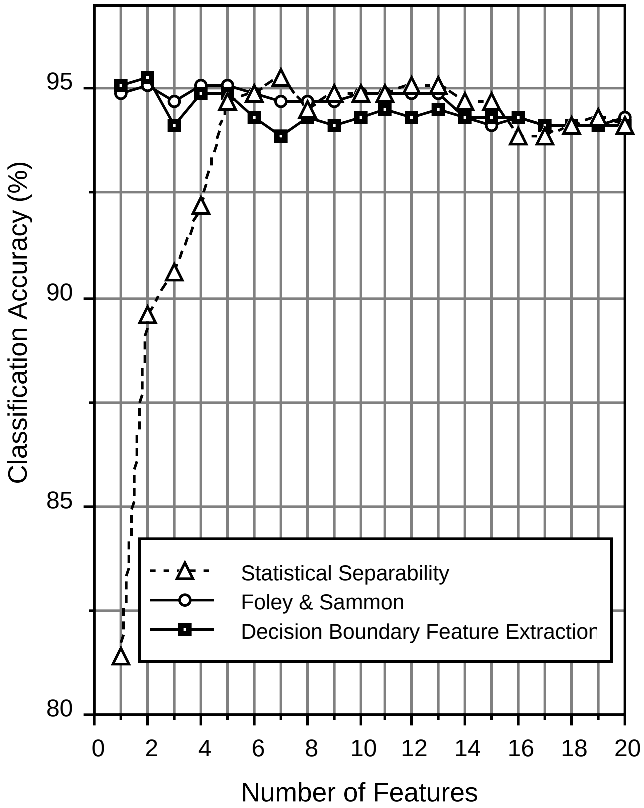

Fig. 14 Performance comparison of Statistical Separability, the Foley & Sammon method, and Decision Boundary Feature Extraction

Table IV shows the eigenvalues of the decision boundary feature matrix along with proportions and accumulations. The eigenvalues are sorted in the decreasing order. The classification accuracies obtained using the corresponding eigenvectors are also showed along with the normalized classification accuracies obtained by dividing the classification accuracies by the classification accuracy obtained using all features. The rank of the decision boundary feature matrix(Σ_{DBFM}) must be decided. Although it is relatively easy to decide the rank for low dimensional generated data, it becomes less obvious for high dimensional real data. One may add eigenvalues until the accumulation exceeds 95% of the total sum and set that number of the eigenvalues as the rank of the Σ_{DBFM}. Defined in this way, the rank of the Σ_{DBFM} would be 5. Alternatively, one may retain the eigenvalues greater than one tenth of the largest eigenvalue. In this way, the rank of the Σ_{DBFM} would be 4. We will discuss more about this problem later.

TABLE IV. Eigenvalues of the Decision Boundary Feature Matrix of the 2 classes along with proportions and accumulations. The classification accuracies are also shown along with the normalized classification accuracies. Ev.:Eigenvalue, Pro. Ev.:Proportion of Eigenvalue, Acc. Ev.: Accumulation of Eigenvalues, Cl. Ac.: Classification Accuracy,
N. Cl. Ac.:Normalized Classification Accuracy(see text).

	Ev.	Pro. Ev. (%)	Acc. Ev. (%)	Cl. Ac. (%)	N. Cl. Ac. (%)
1	0.994	49.6	49.6	93.4	97.9
2	0.547	27.3	77.0	94.3	98.8
3	0.167	8.3	85.3	94.4	99.0
4	0.133	6.6	91.9	95.0	99.6
5	0.066	3.3	95.2	95.1	99.7
6	0.041	2.1	97.3	94.9	99.5
7	0.020	1.0	98.3	94.9	99.5
8	0.012	0.6	98.8	94.8	99.4
9	0.008	0.4	99.2	95.0	99.6
10	0.007	0.3	99.6	95.3	99.9
11	0.005	0.2	99.8	95.3	99.9
12	0.001	0.1	99.9	95.7	100.3
13	0.001	0.0	99.9	95.5	100.1
14	0.001	0.0	100.0	95.4	100.0
15	0.000	0.0	100.0	95.3	99.9
16	0.000	0.0	100.0	95.6	100.2
17	0.000	0.0	100.0	95.5	100.1
18	0.000	0.0	100.0	95.5	100.1
19	0.000	0.0	100.0	95.4	100.0
20	0.000	0.0	100.0	95.4	100.0

In the final test, 4 classes chosen from the data collected at Hand Co. SD. on May 15, 1978. Table V shows the number of samples in each of the 4 classes. Fig. 15 shows the mean graph of the 4 classes. As can be seen, the mean difference is relatively small among some classes. In this test, all data are used for training and test.

Table V. Class Description

SPECIES	DATE	No. of Samples
Winter Wheat	May 15, 1978	223
Native Grass Pas	May 15, 1978	196
Oats	May 15, 1978	163
Unknown Crops	May 15, 1978	253

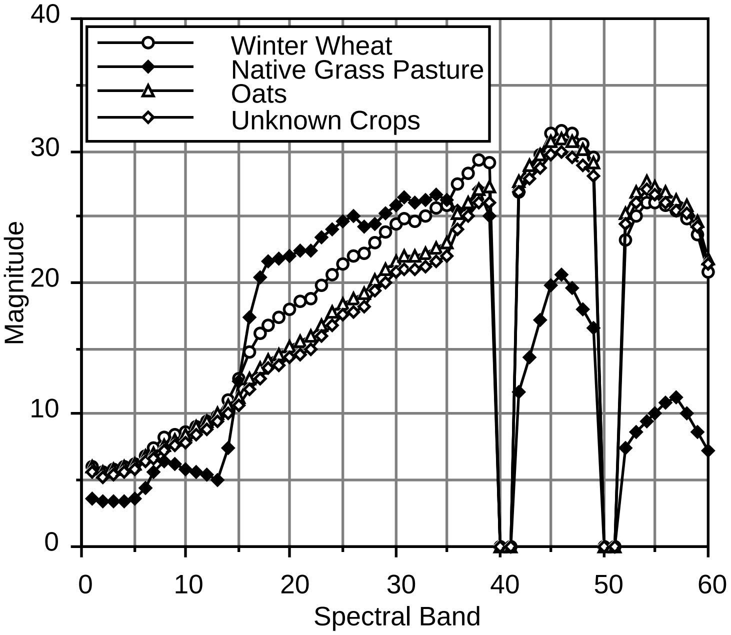

Fig. 15 Mean graph of the 4 classes in Table IV.

Fig. 16 show the performance comparison of the 3 feature selection/extraction algorithms for different numbers of features. The classification accuracy using all features is 88.4%. In this case, CANONICAL ANALYSIS performs less well since class mean differences are relatively small. The performance of DECISION BOUNDARY FEATURE EXTRACTION is much better than those of the other methods. With 8 features, the classification accuracies of DECISION BOUNDARY FEATURE EXTRACTION, CANONICAL ANALYSIS, and STATISTICAL SEPARABILITY are 85.1%, 77.4%, and 76.1%, respectively. DECISION BOUNDARY FEATURE EXTRACTION achieves approximately 87.5% classification accuracy with 11 features while the other methods need 17 features to achieve about the same classification accuracies.

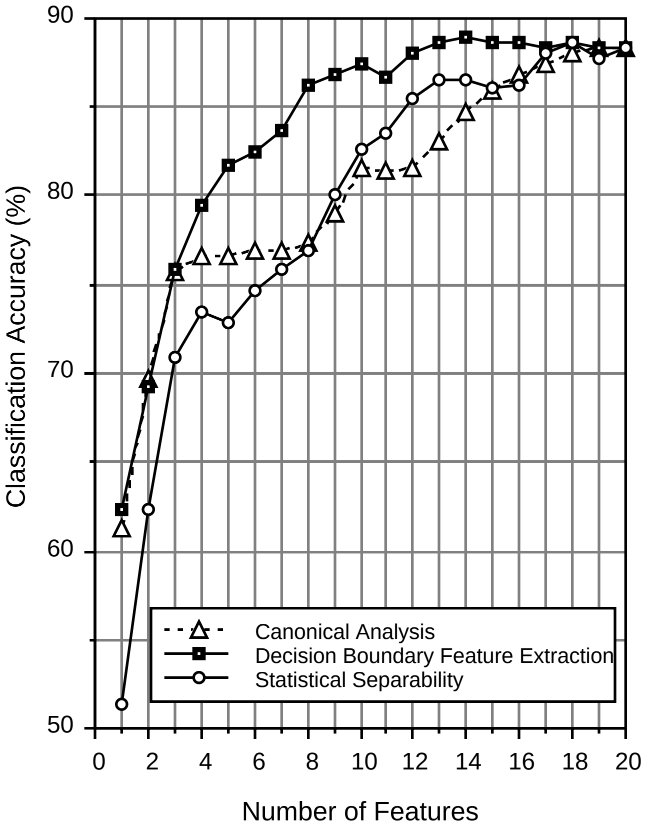

Fig. 16 Performance comparison of Statistical Separability, Canonical Analysis, and Decision Boundary Feature Extraction

Table VI shows the eigenvalues of the decision boundary feature matrix of the 4 classes in Table V along with proportions and accumulations. The classification accuracies obtained using the corresponding eigenvectors are also showed along with the normalized classification accuracies obtained by dividing the classification accuracies by the classification accuracy obtained using all features. Depending on how the threshold is set, the rank of the decision boundary feature matrix could be said to be between 3 to 6. The classification accuracy obtained using all features is 88.4% while the classification accuracies obtained using 3 and 6 features found by DECISION BOUNDARY FEATURE EXTRACTION are 75.8% and 82.5 %, respectively.

TABLE VI . Eigenvalues of the Decision Boundary Feature Matrix of the 4 classes in Table V along with proportions and accumulations. The classification accuracies are also shown along with the normalized classification accuracies. Ev.:Eigenvalue, Pro. Ev.:Proportion of Eigenvalue, Acc. Ev.: Accumulation of Eigenvalues, Cl. Ac.: Classification Accuracy, N. Cl. Ac.:Normalized Classification Accuracy(see text).

	Ev.	Pro. Ev. (%)	Acc. Ev. (%)	Cl. Ac. (%)	N. Cl. Ac. (%)
1	2.956	61.5	61.5	62.3	70.5
2	.917	19.1	80.6	69.3	78.4
3	.317	6.6	87.1	75.8	85.7
4	.193	4.0	91.2	79.5	89.9
5	.157	3.3	94.4	81.7	92.4
6	.109	2.3	96.7	82.5	93.3
7	.066	1.4	98.1	83.7	94.7
8	.042	0.9	99.0	86.3	97.6
9	.029	0.6	99.6	86.8	98.2
10	.009	0.2	99.8	87.4	98.9
11	.007	0.1	99.9	86.7	98.1
12	.002	0.0	100.0	88.0	99.5
13	.002	0.0	100.0	88.6	100.2
14	.000	0.0	100.0	89.0	100.7
15	.000	0.0	100.0	88.6	100.2
16	.000	0.0	100.0	88.6	100.2
17	.000	0.0	100.0	88.4	100.0
18	.000	0.0	100.0	88.7	100.3
19	.000	0.0	100.0	88.4	100.0
20	.000	0.0	100.0	88.4	100.0

Theoretically, the eigenvectors of the decision boundary feature matrix corresponding to non-zero eigenvalues will contribute to improvement of classification accuracy. However, in practice, a threshold must be set to determine the effectiveness of eigenvectors by the corresponding eigenvalues, especially for a high dimensional real data. Fig. 17 shows the relationship between the accumulation of eigenvalues of the decision boundary feature matrix and the normalized classification accuracies obtained by dividing the classification accuracies by the classification accuracy obtained using all features. There is a nearly linear relationship between normalized classification accuracy and accumulation of eigenvalues up to x=95 where x is the accumulation of eigenvalues. As the accumulation of eigenvalues approaches 100 percent, the linear relationship between the normalized classification accuracy and the accumulation of eigenvalues does not hold; care must be taken to set the threshold. More experiments are needed to obtain a better understanding on the relationship between the normalized classification accuracy and the accumulation of eigenvalues.

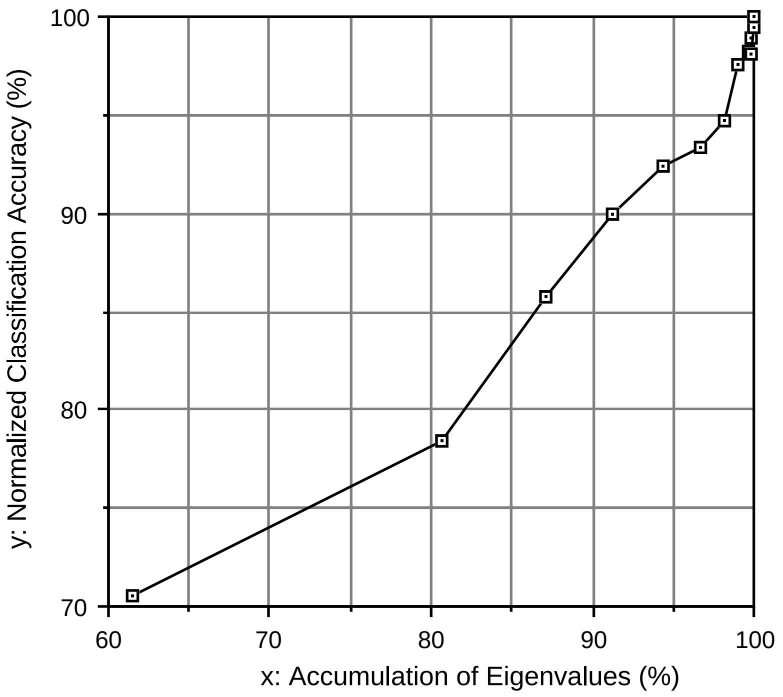

Fig. 17 Relationship between the normalized classification accuracy(see text) and the accumulation of eigenvalues.

VII. CONCLUSION

We have proposed a new approach to feature extraction for classification based on decision boundaries. We defined discriminantly redundant features and discriminant informative features for the sake of feature extraction for classification and showed that the discriminantly redundant features and the discriminantly informative features are related to the decision boundary. By recognizing that normal vectors to the decision boundary are discriminantly informative, the decision boundary feature matrix was defined using the normal vectors. It was shown that the rank of the decision boundary feature matrix is equal to the intrinsic discriminant dimension, and the eigenvectors of the decision boundary feature matrix corresponding to non-zero eigenvalues are discriminantly informative. We then proposed a procedure to calculate empirically the decision boundary feature matrix.

Except for some special cases, the rank of decision boundary feature matrix would be the same as the original dimension. However, it was noted that in many cases only a small portion of the decision boundary is effective in discriminating among pattern classes, and it was shown that it is possible to reduce the number of features by utilizing the effective decision boundary rather than the complete boundary.

The proposed feature extraction algorithm based on the decision boundary has several desirable properties. The performance of the proposed algorithm does not deteriorate even when there is little or no difference in the mean vectors or covariance matrices. In addition, the proposed algorithm predicts the minimum number of features required to achieve the same classification accuracy as in the original space for a given problem. Experiments show that the proposed feature extraction algorithm finds the right feature vectors even in cases where some previous algorithms fail to find them and the performance of the proposed algorithm compares favorably with that of several previous algorithms.

Developments with regard to sensors for Earth observation are moving in the direction of providing much higher dimensional multispectral imagery than is now possible. The HIRIS instrument now under development for the Earth Observing System (EOS), for example, will generate image data in 192 spectral bands simultaneously. In order to analyze data of this type, new techniques for all aspects of data analysis will no doubt be required. The proposed algorithm provides such a new and promising approach to feature extraction for classification of such high dimensional data.

Even though the experiments are conducted using multivariate Gaussian data or assuming a Gaussian distribution, all the developed theorems hold for other distributions or to other decision rules as well. In addition, the proposed algorithm can be also applied for non-parametric classifiers if the decision boundary can be found numerically [16,17].

REFERENCES

[1] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*", John Wiley & Sons, 1973.

[2] J. A. Richards, *Remote Sensing Digital Image Analysis*, Springer-Verlag, 1986.

[3] K. Fukunaga and W. L. G. Koontz, "Application of the Karhunen-Loeve Expansion to Feature Selection and Ordering", IEEE Trans. Computer, Vol. C-19, No. 4, pp 311-318, April 1970.

[4] D. H. Foley and J. W. Sammon, "An Optimal Set of Discriminant Vectors," IEEE Trans. Computer, vol. C-24, No. 3, pp.281-289, Mar. 1975.

[5] D. Kazakos, "On the Optimal Linear Feature," IEEE Trans. Information Theory, vol. IT-24, No. 5, pp.651-652, Sept. 1978.

[6] R. P. Heydorn, "Redundancy in Feature Extraction," IEEE Trans. Computer, pp.1051-1054, Sep. 1971.

[7] P. H. Swain & R. C. King, "Two Effective Feature Selection Criteria for Multispectral Remote Sensing", Proc. First Int. Joint Conf. on Pattern Recognition, 536-540, Nov. 1973.

[8] P. Devijver & J. Kittler, *Pattern Recognition: A Statistical Approach*, Prentice/Hall International, 1982.

[9] W. Malina, "On an Extended Fisher Criterion for Feature selection", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. PAMI-3, No. 5, September 1981.

[10] I. D. Longstaff, "On Extensions to Fisher's Linear Discriminant Function", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. PAMI-9, No. 2, March 1987.

[11] S. D. Morgera and L. Datta, "Toward a Fundamental Theory of Optimal Feature Selection: Part I," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. PAMI-6, No. 5, September 1984.

[12] K. Fukunaga, *Introduction to Statistical Pattern Recognition,* Academic Press, 1972.

[13] C. G. Cullen, *Matrices and Linear Transformation,* Addison Wesley, 1972.

[14] L. L. Biehl, M. E. Bauer, B. F. Robinson, C. S. T. Daughtry, L. F. Silva and D. E. Pitts, "A Crops and Soils Data Base For Scene Radiation Research," Proceedings of the Machine Processing of Remotely Sensed Data Symposium, West Lafayette, IN 1982, pp 169-177.

[15] P. H. Swain & S.M. Davis, *Remote Sensing: The Quantitative Approach*, McGraw–Hill, 1978.

[16] C. Lee & D. A. Landgrebe, "Decision Boundary Feature Selection for Non-Parametric Classifiers," in Proc. SPSE's 44th Annual Conference, pp. 475-478, 1991.

[17] C. Lee & D. A. Landgrebe, "Feature Extraction and Classification Algorithms for High Dimensional Data," PhD Dissertation in School of Electrical Engineering, Purdue University, West Lafayette, Indiana, 1992.