

## Short Paper

---

# A New Classification Approach using Discriminant Functions

AŞKIN DEMIRKOL, ZAFER DEMİR<sup>+</sup> AND EROL EMRE

*Department of Computer Engineering*

<sup>+</sup>*Department of Electrical and Electronic Engineering*

*Sakarya University*

*54187 Sakarya, Turkey*

*E-mail: {askind, zdemir, eemre}@sakarya.edu.tr*

In this study, an approach involving new types of cost functions is given for the construction of discriminant functions. Centers of mass, not specified a priori, around feature vectors are clustered using cost function. Thus, the algorithms yield both the centers of mass and the distinct classes.

**Keywords:** classification, feature vectors, linear discriminant function, Fisher's LDF, dimension reduction

## 1. INTRODUCTION

There are many algorithms for, and many applications of classification and discrimination (grouping of a set of objects into subsets of similar objects where the objects in different subsets are different) in several diverse fields [2-15, 23, 24], ranging from engineering to medicine, to econometrics, etc. Some examples are automatic target recognition (ATR), fault and maintenance-time recognition, optical character recognition (OCR), speech and speaker recognition, etc.

In this study, a new approach and algorithm to the classification problem are described with the goal of finding a single (possibly vector-valued) linear discriminant function. This approach is in terms of some optimal centers of mass for the transformed feature vectors of each class, the transforms being performed via the discriminant functions. As such, it follows the same philosophy which is behind the approaches such as principal component analysis (PCA), Fisher's linear discriminant functions (LDF), and minimum total covariance (MTC) [1-16, 22, 25-28], providing alternatives which extend this work.

Linear discriminant functions (LDF) are often used in pattern recognition to classify a given object or pattern, based on its features, into one of several given classes. For simplicity, consider the discrimination problem for two classes. Let  $x = [x_1, x_2, \dots, x_m]$  be the

---

Received April 28, 2003; revised March 1 and March 29, 2004; accepted May 3, 2004.  
Communicated by H. Y. Mark Liao.

vector representing an object in terms of its  $m$  features. The choice of a small set of features to represent objects is an interesting problem in itself and is not addressed by the theory of linear discriminant functions. The classification task is solved by defining a linear discriminant functions  $g(x)$  defined on the feature vector  $x$

$$g(x) = w^T x + \omega_0, \quad (1)$$

where  $w = [\omega_1, \omega_2, \dots, \omega_m]^T$  is called the weight vector and  $\omega_0$  is called the threshold weight. The object represented by  $x$  is assigned to the class labeled 1 if  $g(x) > 0$ ; otherwise, it is assigned to the class labeled 2. More generally, the discriminant function  $g$  can be written as:

$$g(y) = ay, \quad (2)$$

where  $y = [1, x]^T$  and  $a = [\omega_0, w]$ . If the number of classes is two or more, say  $n$ , then discriminant functions can be defined to distinguish among classes. This approach requires  $n(n-1)/2$  discriminant functions to be defined. The set of classes is denoted by  $\Omega$ , where  $\Omega = \{c_i \mid i = 1, \dots, n\}$  and  $c_i$  refers to the  $i$ -th class [17].

In linear Fisher discriminant analysis, given a set of  $n$   $d$ -dimensional samples  $\{x_1, x_2, \dots, x_n, x_i \in \mathfrak{R}^d\}$  with  $n_1$  samples in class  $C_1$  and  $n_2$  samples in class  $C_2$ . Then the Fisher linear discriminant is given by  $w$ , ( $w \in \mathfrak{R}^d$ ) which maximizes

$$J(w) = \frac{w^T S_B w}{w^T S_W w}, \quad (3)$$

where

$$S_B = (m_1 - m_2)(m_1 - m_2)^T, \quad (4)$$

and

$$S_W = \sum_{i=1,2} \sum_{x \in C_i} (x - m_i)(x - m_i)^T. \quad (5)$$

$S_B$  is known as the between-class scatter matrix,  $S_W$  is called the within-class scatter matrix and  $m_i$  is sample mean of the respective classes defined as

$$m_i = \frac{1}{n_i} \sum_{x \in C_i} x. \quad (6)$$

The reasoning behind maximizing  $J(w)$  as is to look for a direction  $w$  which maximizes the difference between the two projected means in  $S_B$  while minimizing the variance of the individual classes  $S_W$ . Hence samples belonging to two different classes are well separated by projection onto this optimal direction. Furthermore, if the distributions are normal and have equal covariance for the two different classes, then the resulting

linear discriminant function is in the same direction as the Bayes optimal classifier.

In this respect, the problem of constructing classifiers for a small number of samples has been given a solution in [18].

In general, the technique in the study involves two successive optimization stages. The first stage is in terms of either deterministic least squares, or Gauss-Markov (linear unbiased minimum variance) estimation. The second stage is an optimization problem in terms of ratios of two (possibly, matrix-valued) quadratic forms. The solutions of the optimization problems at the second stage are equivalent to obtaining the generalized eigenvalues and eigenvectors of a pair of symmetric positive semidefinite (psd) matrices. In the case of Gauss-Markov estimation, the second stage also involves an additional problem of extremizing the eigenvalues as a function of certain additional parameters, with the prospect of obtaining a better discriminant function as a result.

There are many algorithms to solve the first stage. The first stage defines the structure of the linear discriminant function with some free parameters which, in turn, are determined at the second stage via optimizing a cost function of the type mentioned above. Throughout, linear functions will be considered for use as discriminant functions for classification. However, this is really not a substantial loss of generality, because, as shown next, a large class of nonlinear functions for classification can be obtained directly using the techniques established for linear classifiers. Such functions increase the dimension of the feature vectors in a straightforward way.

A typical paradigm for the type of classification problems addressed here consists of three main stages.

- (i) **Representation:** At this stage, physical objects are represented quantitatively (i.e., as numbers), such as pixel values of images, or amplitude measurements for signals, etc. [1-14].
- (ii) **Feature Selection:** As the number of values in the quantitative representation can be very large, and not all the details present in the representation may be necessary for the purpose of classification, typically, a certain set (much smaller in number than the representation) of functions on the representative values are evaluated, associating with each object a vector of numbers, called the feature vector. These functions are chosen to represent the most relevant properties of the objects with regard to the particular classification (e.g., mean, variance, skewness, kurtosis, spectral quantities). There is no application independent and universally accepted set of functions to be used for selecting the feature vector. The features chosen usually depend very much on the particular set of objects, and the purpose of classification. [1-14, 18, 19].
- (iii) **Classification and Discrimination:** Using some algorithm (such as the ones developed here), construction of a discriminant function will separate feature vectors (hence, the objects) into possible distinct classes.

The *classification and discrimination stage* can be in two different forms [1-14, 20]:

The first form is supervised learning where the classes have been specified already, and a subset of the feature vectors, called a training set, is given for use in constructing a

mathematical rule which can determine the class of any given feature vector (referred to an abstraction, generalization, or induction). Such a rule involves a function called a discriminant function. This paper primarily considers the first case.

The second form is unsupervised learning, where there is no a priori information about any type of predetermined classes. The problem is to detect structures, commonality and differences among the given feature vectors, and then separate them accordingly into different subsets (the classes).

The goal of this work is examined in detail in sections 2 and 3.

## 2. A NEW TECHNIQUE TO OBTAIN A LINEAR DISCRIMINANT FUNCTION AND REDUCTION OF THE DIMENSION OF THE FEATURE SPACE

Deterministic least squares will be used as the first optimization stage. As it will be clear later in this section, with the approach described here, the two dimension reduction techniques that are derived here both yield a possibly vector-valued linear discriminant function. However, for ease of exposition, first the techniques will be explained in terms of obtaining a scalar-valued linear discriminant function. Then the extensions to dimension reduction and vector-valued discriminants will be easy to describe.

Suppose that there are  $c$  classes  $C_1, \dots, C_c$ . Let  $N_i, 1 \leq i \leq c$  be the number of given feature vectors of class  $C_i$ . Let  $h$  be a  $p$ -dimensional row vector denoting a feature vector.

Define  $p$  and  $m$ . We wish to find a  $p \times m$  matrix  $\hat{X}$  (the discriminant), and  $m$ -dimensional row vectors  $y_0, y_1, \dots, y_c$  (the centers of masses) such that (using the Euclidean distances as the measure of distances),

- (a)  $h\hat{X} - y_i$  is small, and
- (b)  $h\hat{X} - y_j$  is large,

for all  $h$  in  $C_i, j \neq i$  and  $1 \leq i, j \leq c$ .

Define

$$Y = \begin{pmatrix} y_0 \\ y_1 \\ \cdot \\ \cdot \\ y_c \end{pmatrix} = (Y_1, Y_2, \dots, Y_m) = (e_{ij}), \quad (7)$$

where  $Y_k, 1 \leq k \leq m$ , is the  $k$ -th column of  $Y$ , and  $e_{ij}$  is the  $ij$ -th entry. Let  $\hat{X}_k$  be the  $k$ -th column of  $\hat{X}$ .

Let  $h_1, h_2, \dots, h_{N_1}$  be the given feature vectors in  $C_1$ , and let  $h_{N_1+1}, \dots, h_{N_1+N_2}$  be the given feature vectors in  $C_2$ , i.e.,  $h_i$ 's form the training set. Define the matrices  $H_1, H_2, \dots$ ,

$H$  as follows.  $H_1$  is the matrix whose rows are  $h_1, \dots, h_{N_1}$ .  $H_2$  is the matrix whose rows are  $h_{N_1+1}, \dots, h_{N_1+N_2}, \dots$ .  $H$  is obtained by stacking  $H_1, H_2, \dots$ , from the top to the bottom. Let  $N = \sum_{i=1}^c N_i$ .

First the least squares technique is applied to force  $h_i \hat{X}$  closer to its center of mass,  $y_i$ ,  $1 \leq i \leq N$ , i.e., to enforce the requirement (a) above.

Thus, the first step is to find a matrix  $\hat{X}$  such that

$$E = \sum_{i=1}^N \|h_i \hat{X} - \tilde{y}_i\|^2 = \sum_{i=1}^N \sum_{j=1}^m (h_i \hat{X}_j - \tilde{e}_{ij})^2, \tag{8}$$

is minimum, where  $1 \leq i \leq c$  and  $i = l$  iff  $h_i \in C_l$ , i.e.,

$$\tilde{y}_i = y_l, \text{ and } \tilde{e}_{ij} = e_{lj} \text{ iff,} \tag{9}$$

$$h_i \in C_l,$$

$\| \cdot \|$  denotes the Euclidean norm for vectors. Exchanging the order of summation above, we can write

$$E = \sum_{j=1}^m \|H(\hat{X})_j - MY_j\|^2, \tag{10}$$

where

$$M = \text{blockdiag}\{\hat{1}_l\}_{l=1}^c, \tag{11}$$

and  $\hat{1}_l$  is the column vector of size  $N_l$  all of whose entries are  $l$ 's. Clearly, minimization of each term separately, is equivalent to the minimization of  $E$ . This requires that

$$\hat{X}_j = H^+ MY_j, \quad 1 \leq j \leq m, \tag{12}$$

where  $R^+$  denotes the unique (Moore-Penrose) **Pseudoinverse** [1-14] of any matrix  $R$ . We should note that any generalized inverse would work instead of Moore-Penrose inverse, but there are better techniques to obtain this inverse. In particular if  $H$  has linearly independent columns, as it will in many applications, then

$$H^+ = (H^T H)^{-1} H^T.$$

Hence

$$\hat{X} = H^+ MY, \tag{13}$$

and

$$E = \left\| H\hat{X} - MY \right\|^2, \quad (14)$$

where, now the norm above is the Euclidean norm of the matrix

$$HH^+MY - MY = (HH^+ - I)MY = P_0MY. \quad (15)$$

Note that  $P_0$  above is the orthogonal projection on  $R(H)^\perp$  (the orthogonal complement of the range space of  $H$ ).  $E$  in Eq. (14) can be written as

$$E = \text{tr}(Y^T M^T P_0 M Y) = \text{tr}(Y^T B Y). \quad (16)$$

Thus, Eq. (16) represents the criterion for the transformed (via  $\hat{X}$ ) class members to be close to their centers of masses. While this is a very desirable property, it is not necessarily sufficient by itself. It is essential that the rows of  $Y$  be distinct, and the transformed feature vectors be sufficiently far from the centers of masses of other classes. No  $Y$  which satisfies only one of these two properties at the expense of the other is an acceptable solution.

Before considering the criterion for the choice of  $Y$ , the expression for the latter (distance to other centers of masses) will now be given.

Total distance from the members of class  $i$  to the other centers of masses can be derived (similar to the above derivation) to be

$$D_i^2 = \text{tr}(Y^T M_i^T P_i M_i Y) = \text{tr}(Y^T A_i Y), \quad (17)$$

where  $M_i$  is the matrix with  $c - 1$  row blocks of size  $N_i \times c$  obtained from the matrix  $\text{blockdiag}\{\hat{1}_l\}_{l=1}^c$  (where  $\hat{1}_l$  is a column vector of size  $N_l$  all of whose entries are 1's), by deleting its  $i$ -th row block.

$$P_i = D_i^T D_i, \quad (18)$$

$$D_i = (\tilde{H}_i H^+ - I), \quad (19)$$

where  $\tilde{H}_i$  is the matrix obtained by stacking  $c - 1$  copies of  $H_i$  on top of each other. Thus, the total distance squared is

$$D^2 = \sum_{i=1}^c \text{tr}(Y^T A_i Y) \quad (20)$$

$$= \text{tr}(Y^T A Y). \quad (21)$$

It is seen that the matrix  $Y$  is so far unspecified. The above expressions are true for any choice of  $Y$ . Next, we will exploit this freedom to further enforce requirement (a), while also enforcing requirement (b) above.

This can be approached by maximizing (with respect to  $Y$ ) a certain cost function  $f(x, y)$  where one can substitute

$$x = m(Y^T A Y), \quad (22)$$

$$y = m(Y^T B Y), \quad (23)$$

which is strictly increasing in the first argument, and strictly decreasing in the second argument (for each fixed value of the other argument for which  $f$  is defined), and where  $m$  is some measure of magnitude of the matrix which is its argument.

Clearly, one can conceive of many such functions  $f$ . Of course, a major consideration in this choice is the numerical solvability of the resulting optimization problem. A simple candidate function  $f$  is the ratio of the two arguments,  $f = x/y$  with  $m$  chosen as the determinant. This is used in Fisher's LDF. However, if there exists a matrix  $Y$  which makes the denominator zero (as is the case when  $H$  has rank less than or equal to its number of columns), then even though there may be more than one such matrix  $Y$ , the cost function does not take into consideration the maximization of the numerator at all.

The function that we utilize is a slight modification of the above function, namely,

$$f = x/\bar{y}, \quad (24)$$

where

$$\bar{y} = m(Y^T Y + Y^T B Y). \quad (25)$$

We choose  $m$  to be the determinant ( $m = \det$ ).

Clearly, now even if  $Y$  is such that  $y = \det(Y^T B Y)$  is zero,  $I + B$  being positive definite, still the term

$$x/\det(Y^T B Y + Y^T Y), \quad (26)$$

will be maximized among such  $Y$ 's.

When  $Y$  is a column vector and all quantities above are scalar (note that for a scalar, its trace and determinant are equal to itself), a scalar-valued linear discriminant function is being sought, and also using Eqs. (16) and (17),  $f$  becomes

$$f = D^2 / (Y^T Y + E^2). \quad (27)$$

Then an optimal solution is given by a generalized eigenvector of the pair of matrices  $(I + B, A)$  (where  $I$  is the identity matrix), corresponding to the largest generalized eigenvalue. In this case, the choice of  $m$  either as determinant or trace leads to the same result. Finding a scalar-valued linear discriminant function is equivalent to projecting the feature vectors into a line, a one dimensional subspace of the feature space. However, in some cases, especially when there are more than two classes, it may be too much to ask for a single linear discriminant function to perform satisfactorily. Then, a more promising approach is to project the feature vectors to a lower dimensional subspace, but greater than one, while also finding a vector center of mass for the transforms of the feature vectors in each class to cluster around. Then, clearly this corresponds to the case of  $m \geq 2$ .

It is also important to note that for  $m \geq 2$ , the columns of  $Y$  must be linearly independent. Otherwise, introduction of a new column in  $Y$  does not lead to different rows

from the possibly identical rows of  $Y$  (centers of masses). This is clearly unacceptable as a solution.

In this case,  $m = \text{determinant}$  as in Fisher's LDF again leads to the solution  $Y$  whose  $m$  generalized eigenvectors correspond to the  $m$  largest generalized eigenvalues. This also provides the solution with  $m = \text{trace}$  provided that  $Y$  is constrained so that the denominator is the identity matrix. If no condition is imposed on  $Y$ , the optimal solution is a matrix with all columns being zero except for one which is a generalized eigenvector corresponding to the largest generalized eigenvalue, which is clearly unacceptable.

### 3. CONCLUSIONS

In this study, a new approach is given for construction of more effective discriminant functions. This approach extends primarily the work of Fisher [1, 18, 21]. The discriminant functions can be obtained via solving a classical generalized eigenvalue – eigenvector problem (whose solution is well-known). The obtained discriminant functions map members of each class closer together. Simultaneously, they map members of different classes further from each other. It is expected that these discriminant functions will serve better to separate distinct classes while clustering members of the same class closer to each other. Therefore the discriminant functions appear more reasonable for automatic target recognition, optical character recognition, face recognition, speech and speaker recognition.

### REFERENCES

1. R. Duda and P. Hart, *Pattern Classification and Scene Analysis*, John Wiley and Sons, New York, 1973.
2. D. Reisfeld, H. Wolfson, and Y. Yeshurun, "Context free attentional operators: the generalized symmetry transform," *International Journal of Computer Vision, Special Issue on Qualitative Vision*, Vol. 14, 1995, pp. 119-130.
3. B. Kumar, D. Casasent, and H. Murakami, "Principal component imagery for statistical pattern recognition correlators," *Optical Engineering*, Vol. 21, 1982, pp. 43-47.
4. S. S. Wilks, *Mathematical Statistics*, John Wiley and Sons, New York, 1962.
5. B. Moghaddam, "Probabilistic visual learning for object detection," in *Proceedings of 5th International Conference on Computer Vision*, 1995, pp. 786-793.
6. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley and Sons, New York, 1994.
7. P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, 1997, pp. 711-720.
8. Y. Cui, D. Swets, and J. Weng, "Learning-based hand sign recognition using SHOSLIF- M," *International Conference on Computer Vision*, 1995, pp. 631-636.
9. R. J. Woodham, "Analyzing images of curved surfaces," *Journal of Artificial Intelligence*, Vol. 17, 1981, pp. 117-140.
10. M. J. Karson, *Multivariate Statistical Methods*, The Iowa State University Press,

- 1982.
11. S. Watanabe, *Pattern Recognition: Human and Mechanical*, John Wiley and Sons, New York, 1985.
  12. D. J. Hand, *Construction and Assessment of Classification Rules*, John Wiley and Sons, New York, 1997.
  13. S. J. Roberts, D. Husmeier, and L. Rezek, "Bayesian approaches to Gaussian mixture modeling," *IEEE Transactions on Pattern Recognition and Machine Intelligence*, Vol. 20, 1998, pp. 1133-1142.
  14. S. Raudys, "On dimensionality, sample size, and classification error of nonparametric linear classification algorithms," *IEEE Transactions on Pattern Recognition and Machine Intelligence*, Vol. 19, 1997, pp. 669-671.
  15. M. Westphal, T. Schultz, and A. Waibel, "Linear discriminant – a new criterion for speaker normalization," in *Proceedings of International Conference on Spoken Language (ICSLP '98)*, Vol. 3, 1998, pp. 827-830.
  16. D. B. Heras, J. C. Cabaleiro, V. B. Perez, P. Costas, and F. F. Rivera, "Principal component analysis on vector computers," in *Proceedings of Vector and Parallel Processing (VECPAR)*, 1996, pp. 416-428.
  17. Z. Q. Hong and J. Y. Yang, "Optimal discriminant plan for a small number of samples and design method of classifier on the plan," *Pattern Recognition*, Vol. 24, 1991, pp. 317-324.
  18. F. Model, P. Adorjan, A. Olek, and C. Piepenbrock, "Feature selection for DNA methylation based cancer classification," *Bioinformatics Discovery Note*, Vol. 1, 2001, pp. 1-8.
  19. A. Talukder and D. Casaent, "Nonlinear features for classification and pose estimation of machined parts from single views," in *Proceedings of Society of Photo Optical Instrumentation Engineers (SPIE)*, Vol. 3522, 1998, pp. 16-27.
  20. S. A. Billings and K. L. Lee, "Nonlinear Fisher discriminant analysis using a minimum squared error cost function and the orthogonal least squares algorithm," *Neural Networks*, Vol. 15, 2002, pp. 263-270.
  21. A. Biem, S. Katagiri, and B. H. Juang, "Pattern recognition using discriminative feature extraction," *IEEE Transactions on Signal Processing*, Vol. 45, 1997, pp. 500-504.
  22. W. Zhao, R. Chellappa, and A. Krishaswamy, "Discriminant analysis of principal components for face recognition," in *Proceedings of 3rd International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 336-341.
  23. J. Yang, H. Yu, and W. Kunz, "An efficient LDA algorithm for face recognition," in *Proceedings of 6th International Conference on Control, Automatic, Robotics and Vision (ICARCV 2000)*, 2000.
  24. W. Chou, "Discriminant-function-based minimum recognition error rate pattern recognition approach to speech recognition," in *Proceedings of the IEEE*, Vol. 88, 2000, pp. 1201-1223.
  25. R. P. W. Duin and R. H. Umbach, "Multiclass linear dimension reduction by weighted pairwise Fisher criteria," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, 2001, pp. 762-766.
  26. H. Ujiie, S. Omachi, and H. Aso, "A discriminant function considering normality improvement of the distribution," in *Proceedings of 16th International Conference*

*on Pattern Recognition (ICPR 2002)*, Vol. 2, 2002, pp. 224-227.

27. J. Yang, D. Zhang, A. F. Frangi, and J. Y. Yang, "Two dimensional PCA: a new approach to appearance based face representation and recognition," *IEEE Transactions on Pattern analysis and Machine Intelligence*, Vol. 26, 2004, pp. 131-137.

**Aşkin Demirkol** graduated from Yildiz Technical University, Electrical Engineering Department in 1988. He received M.S. degree from Istanbul University and Ph.D. degree from Trakya University. He has currently been working for Computer Engineering Department of Sakarya University. His areas of interest include radar signal processing, imaging and pattern recognition. He has been temporarily keeping his research in the Electrical and Computer Engineering Department at University of Missouri, Rolla, U.S.A.

**Zafer Demir** was born in Kutahya, Turkey in 1959. He graduated from A.D.M.M.A Electrical Engineering Department in 1982, Ankara, Turkey. He received M.S. degree from Gazi University and Ph.D. degree from Kocaeli University. He has currently been working for Electrical and Electronic Engineering Department of Sakarya University. His current research interests include, power systems, expert systems, and pattern recognition.

**Erol Emre** received the B.S.E.E. and M.S.E.E. degrees from Middle East Technical University (METU), Ankara, Turkey, in 1973 and 1974, respectively. He received his Ph.D. in Electrical Engineering from the University of Southern California (USC), Los Angeles, in 1976. His present interests are mathematical system theory, system identification/signal processing, pattern recognition, and multitarget tracking/sensor fusion.