C. B. Chittineni*
Lockheed Electronics Company, Inc.
Houston, TX 77058

Fisher Classifier and its Probability of Error Estimation

Computationally efficient expressions are derived for estimating the probability of error using the leave-one-out method.

Introduction

Recently, there has been considerable interest in the development of techniques for the classification of imagery data (such as the data acquired by the Landsat series of satellites) for identifying and inventorying natural resources, monitoring crop conditions, detecting changes in natural and man-made objects, etc. Supervised or nonsupervised classification techniques can be used for the classification of imagery data. Both of these approaches require the availability of class labels for the training patterns. For remote sensing imagery, it can be difficult and expensive to obtain labels for the training patterns. (For example, in the classification of remotely sensed agricultural imagery data, the labels for the training patterns are provided by an analyst interpreter by examining imagery films and using some other information, such as crop growth stage models historic information, etc.)

Usually, as the number of parameters to be estimated increases, the required number of labeled patterns also increases. There is considerable interest in the use of linear classifiers for imagery data classification because they are simple and because fewer parameters need to be estimated than, for

Abstract: *The Fisher classifier and the problem of estimating its probability of error are considered. Computationally efficient expressions are derived for estimating the probability of error using the leave-one-out method. The optimal threshold for the classification of patterns projected onto Fisher's direction is derived. A simple generalization of the Fisher classifier to multiple classes is presented. Furthermore, computational expressions are developed for estimating the probability of error of the multiclass Fisher classifier.*

example, with maximum likelihood clustering. In many cases, it is required to estimate the probabilities of classification errors in addition to designing the classifier. (For example, to correct for the bias introduced by the classifier in the estimation of proportion of class of interest in remotely sensed agricultural imagery data, the probabilities of classification errors are estimated using a separate set of labeled patterns called Type II dots.) Because acquiring labels for the patterns is expensive, the available labeled samples should be effectively used both for designing the classifier and for estimating the probabilities of classification errors.

The leave-one-out method (Lachenbroch and Mickey, 1968) is proposed in the literature as an effective way of estimating the probability of error from the training samples. The method is as follows: If there is a total of N-labeled patterns, leave out one pattern, design the classifier on remaining $(N - 1)$ patterns, and test on the pattern that is left out. Repeat this process N times, every time leaving a different pattern, and then estimate the probability of error as an average of these errors. Use of this method, however, requires N classifiers to be designed. Misra (1979) presented simulation results using remote sensing data in estimating the Fisher error probability with the leave-one-out method. Chitti-

* Now with the Research & Development Department, Conoco Inc., P.O. Box 1267, Ponca City, OK 74601.

Photogrammetric Engineering and Remote Sensing,
Vol. 48, No. 9, September 1982, pp. 1445-1453.

0099-1112/82/4809-1445$02.25/0

neni (1977) developed a computational technique based on eigen perturbation theory for estimating the probability of error of the Fisher classifier using the leave-groups-out method.

This paper considers the Fisher classifier (Fisher, 1963; Chittineni, 1972). The Fisher classifier is one of the most widely used linear classifiers. Computational expressions are developed based on matrix theory for estimating the probability of error of the Fisher classifier using the leave-one-out method. This paper is organized as follows:

The Fisher classifier for the two-class case is first presented. Computational expressions for using the leave-one-out method for estimating Fisher's error probability are then developed. The effect of the Fisher threshold is discussed and expressions for obtaining the optimal threshold by minimizing the probability of error are presented. In the next section, a simple generalization of the Fisher classifier to multiple classes is presented. Finally, computationally efficient expressions for the estimation of multicategory Fisher error using the leave-one-out method are developed. Some matrix relations used in the paper are derived in Appendix A. Detailed derivations of the optimal threshold are presented in Appendix B.

Fisher Classifier

The Fisher classifier is a linear classifier that uses a direction, $\mathbf{W}$, for the discriminant function

$$g(\mathbf{X}) = \mathbf{W}^T\mathbf{X} - t \tag{1}$$

so that, when the training patterns are projected onto this direction, the intraclass patterns are clustered and the interclass patterns are separated to the extent possible, as depicted in Figure 1.

Let $X_k^i \epsilon \omega_i$, $k = 1, 2, \ldots, N_i$, $i = 1, 2$ be the training pattern set. The unbiased estimates of means, $\hat{\mathbf{m}}_i$, and covariance matrices, $\hat{\mathbf{\Sigma}}_i$, of the patterns in the classes, ω_i, are given by the following:

$$\hat{\mathbf{m}}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbf{X}_j^i \tag{2}$$

$$\hat{\mathbf{\Sigma}}_i = \frac{1}{(N_i - 1)} \sum_{j=1}^{N_i} (\mathbf{X}_j^i - \hat{\mathbf{m}}_i)(\mathbf{X}_j^i - \hat{\mathbf{m}}_i)^T. \tag{3}$$

The Fisher classifier chooses the weight vector, $\mathbf{W}$, such that, when the patterns are projected onto it, the interclass distances are maximized and the intraclass distances are minimized. A criterion, β, which can be used to obtain the weight vector, $\mathbf{W}$, can be written as follows:

$$\beta = \frac{[\mathbf{W}^T(\hat{\mathbf{m}}_1 - \hat{\mathbf{m}}_2)]^2}{\mathbf{W}^T\hat{\mathbf{S}}_W\mathbf{W}} \tag{4}$$

where

$$\hat{\mathbf{S}}_W = \hat{\mathbf{\Sigma}}_1 + \hat{\mathbf{\Sigma}}_2. \tag{5}$$

The weight vector, $\mathbf{W}$, which maximizes β, can easily be shown to be

$$\mathbf{W} = (\hat{\mathbf{S}}_W)^{-1}(\hat{\mathbf{m}}_1 - \hat{\mathbf{m}}_2). \tag{6}$$

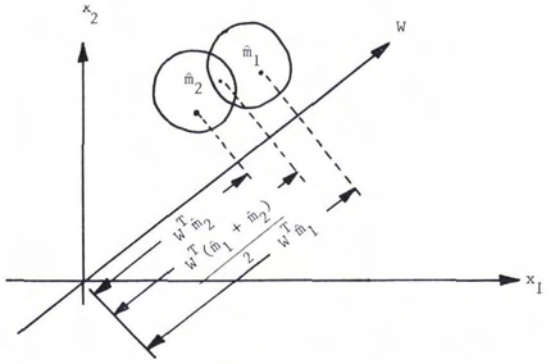

Fig. 1. Fisher's weight vector and threshold.

The Fisher threshold, t, is chosen as

$$t = \mathbf{W}^{\mathrm{T}} \frac{(\hat{\mathbf{m}}_1 + \hat{\mathbf{m}}_2)}{2} . \tag{7}$$

The direction, $\mathbf{W}$, and the threshold, t, are illustrated in Figure 1. Fisher's decision rule is as follows:

$$\text{Decide } \mathbf{X} \epsilon \omega_1 \text{ if } g(\mathbf{X}) \geq 0 \tag{8}$$

$$\text{Decide } \mathbf{X} \epsilon \omega_2 \text{ if } g(\mathbf{X}) < 0. \tag{9}$$

Fisher classifications depend on the direction, $\mathbf{W}$, and the threshold, t. Later in this paper, the effect of Fisher threshold on classifications is discussed, and expressions are presented for obtaining the optimal threshold, t, in the Fisher direction, $\mathbf{W}$, by minimizing the probability of error.

Recursive Relations for the Fisher Weight Vector and Threshold

In this section, computational expressions are developed for using the leave-one-out method with the Fisher classifier described in the previous section. The cases in which a pattern $\mathbf{X}_k^1$ from class ω_1 is left out and in which a pattern from class ω_2 is left out can be treated identically.

Let a pattern $\mathbf{X}_k^1$ from class ω_1 be left out and the patterns from class ω_2 remain. The means $\hat{\mathbf{m}}_i$, $i = 1, 2$, and the covariance matrix $\hat{\mathbf{\Sigma}}_2$ are defined as in Equations 2 and 3. Define the matrix $\hat{\mathbf{\Sigma}}_1$ of the total pattern set from class ω_1 as

$$\hat{\mathbf{\Sigma}}_1 = \frac{1}{(N_1 - 2)} \sum_{j=1}^{N_1} (\mathbf{X}_j^1 - \hat{\mathbf{m}}_1)(\mathbf{X}_j^1 - \hat{\mathbf{m}}_1)^{\mathrm{T}}. \tag{10}$$

Let

$$\hat{\mathbf{S}}_{\mathbf{W}} = \hat{\mathbf{\Sigma}}_1 + \hat{\mathbf{\Sigma}}_2. \tag{11}$$

Note that $\hat{\mathbf{\Sigma}}_1$ is defined differently from the usual unbiased estimate for covariance matrices. The $\hat{\mathbf{\Sigma}}_1$ is defined this way to allow the use of the Bartlett matrix inversion relation (Equations A-5 and A-6) in obtaining computationally efficient recursive relations for the Fisher weight vector and threshold. With this definition, then, for estimating the Fisher probability of error, the matrix inversion needs to be done only twice, once when patterns from class ω_1 are left out and again when patterns from class ω_2 are left out. Now compute $\mathbf{W}$ and t as

$$\mathbf{W} = \hat{\mathbf{S}}_{\mathbf{W}}^{-1} (\hat{\mathbf{m}}_1 - \hat{\mathbf{m}}_2) \tag{12}$$

and

$$t = \mathbf{W}^{\mathrm{T}} \frac{(\hat{\mathbf{m}}_1 + \hat{\mathbf{m}}_2)}{2} . \tag{13}$$

When a pattern $\mathbf{X}_k^1$ from class ω_1 is left out, the unbiased estimates of the mean $\hat{\mathbf{m}}_{1k}$ and the covariance matrix $\hat{\mathbf{\Sigma}}_{1k}$ of the patterns in class ω_1 are given by the following:

$$\hat{\mathbf{m}}_{1k} = \frac{1}{(N_1 - 1)} \sum_{\substack{j=1 \\ \neq k}}^{N_1} \mathbf{X}_j^1 \tag{14}$$

and

$$\hat{\mathbf{\Sigma}}_{1k} = \frac{1}{(N_1 - 2)} \sum_{\substack{j=1 \\ \neq k}}^{N_1} (\mathbf{X}_j^1 - \hat{\mathbf{m}}_{1k})(\mathbf{X}_j^1 - \hat{\mathbf{m}}_{1k})^{\mathrm{T}}. \tag{15}$$

Let $\mathbf{S}_{\mathbf{W}1k} = \hat{\mathbf{\Sigma}}_{1k} + \hat{\mathbf{\Sigma}}_2$. Then the Fisher weight vector $\mathbf{W}_{1k}$ and threshold t_{1k}, when a pattern $\mathbf{X}_k^1$ from class ω_1 is left out, are given by

$$\mathbf{W}_{1k} = \hat{\mathbf{S}}_{\mathbf{W}1k}^{-1} (\hat{\mathbf{m}}_{1k} - \hat{\mathbf{m}}_2) \tag{16}$$

$$t_{1k} = \frac{\mathbf{W}_{1k}^{\mathrm{T}} (\hat{\mathbf{m}}_{1k} + \hat{\mathbf{m}}_2)}{2} . \tag{17}$$

Expressions are now developed for the computation of $\mathbf{W}_{1k}$ and t_{1k} in terms of $\mathbf{W}$ and t. The relationships between $\mathbf{m}_{1k}$, $\hat{\boldsymbol{\Sigma}}_{1k}$, $\hat{\mathbf{S}}_{\mathbf{W}1k}$, and $\hat{\mathbf{m}}_1$, $\hat{\boldsymbol{\Sigma}}_1$, and $\hat{\mathbf{S}}_{\mathbf{W}}$ can be shown to be as follows (see Appendix A):

$$\hat{\mathbf{m}}_{1k} = \hat{\mathbf{m}}_1 - \frac{1}{(N_1 - 1)} (\mathbf{X}_k^1 - \hat{\mathbf{m}}_1) \tag{18}$$

$$\hat{\boldsymbol{\Sigma}}_{1k} = \hat{\boldsymbol{\Sigma}}_1 - \frac{N_1}{(N_1 - 1)(N_1 - 2)} (\mathbf{X}_k^1 - \hat{\mathbf{m}}_1)(\mathbf{X}_k^1 - \hat{\mathbf{m}}_1)^{\mathrm{T}} \tag{19}$$

$$\hat{\mathbf{S}}_{\mathbf{W}1k} = \hat{\mathbf{S}}_{\mathbf{W}} - \frac{N_1}{(N_1 - 1)(N_1 - 2)} (\mathbf{X}_k^1 - \hat{\mathbf{m}}_1)(\mathbf{X}_k^1 - \hat{\mathbf{m}}_1)^{\mathrm{T}} . \tag{20}$$

From Equation 18, one obtains

$$(\hat{\mathbf{m}}_{1k} - \hat{\mathbf{m}}_2) = (\hat{\mathbf{m}}_1 - \hat{\mathbf{m}}_2) - \frac{1}{(N_1 - 1)} (\mathbf{X}_k^1 - \hat{\mathbf{m}}_1). \tag{21}$$

From Equation 20, one obtains (Appendix A)

$$\hat{\mathbf{S}}_{\mathbf{W}1k}^{-1} = \hat{\mathbf{S}}_{\mathbf{W}}^{-1} + \frac{\alpha \hat{\mathbf{S}}_{\mathbf{W}}^{-1} (\mathbf{X}_k^1 - \hat{\mathbf{m}}_1)(\mathbf{X}_k^1 - \hat{\mathbf{m}}_1)^{\mathrm{T}} \hat{\mathbf{S}}_{\mathbf{W}}^{-1}}{1 - \alpha(\mathbf{X}_k^1 - \hat{\mathbf{m}}_1)^{\mathrm{T}} \hat{\mathbf{S}}_{\mathbf{W}}^{-1} (\mathbf{X}_k^1 - \hat{\mathbf{m}}_1)} \tag{22}$$

where

$$\alpha = \frac{N_1}{(N_1 - 1)(N_1 - 2)} . \tag{23}$$

Let

$$\mathbf{Y}(\mathbf{X}_k^1) = \hat{\mathbf{S}}_{\mathbf{W}}^{-1} (\mathbf{X}_k^1 - \hat{\mathbf{m}}_1) \tag{24}$$

$$\beta(\mathbf{X}_k^1) = (\mathbf{X}_k^1 - \hat{\mathbf{m}}_1)^{\mathrm{T}} \hat{\mathbf{S}}_{\mathbf{W}}^{-1} (\mathbf{X}_k^1 - \hat{\mathbf{m}}_1) \tag{25}$$

$$\nu(\mathbf{X}_k^1) = 1 - \alpha\beta(\mathbf{X}_k^1) \tag{26}$$

$$\hat{\mathbf{m}} = \frac{(\hat{\mathbf{m}}_1 + \hat{\mathbf{m}}_2)}{2} \tag{27}$$

$$Z(\mathbf{X}_k^1) = \mathbf{Y}^{\mathrm{T}}(\mathbf{X}_k^1)(\hat{\mathbf{m}}_1 - \hat{\mathbf{m}}_2). \tag{28}$$

Using the definitions of Equations 23 to 28, one obtains the following:

$$\begin{aligned} \mathbf{W}_{1k} &= \hat{\mathbf{S}}_{\mathbf{W}1k}^{-1} (\hat{\mathbf{m}}_{1k} - \hat{\mathbf{m}}_2) = \left[\hat{\mathbf{S}}_{\mathbf{W}}^{-1} + \frac{\alpha \mathbf{Y}(\mathbf{X}_k^1)\, \mathbf{Y}^{\mathrm{T}}(\mathbf{X}_k^1)}{\nu(\mathbf{X}_k^1)} \right] \left[(\hat{\mathbf{m}}_1 - \hat{\mathbf{m}}_2) - \frac{1}{(N_1 - 1)} (\mathbf{X}_k^1 - \hat{\mathbf{m}}_1) \right] \\ &= \mathbf{W} - \frac{1}{(N_1 - 1)\, \nu(\mathbf{X}_k^1)} \mathbf{Y}(\mathbf{X}_k^1) + \frac{\alpha Z(\mathbf{X}_k^1)}{\nu(\mathbf{X}_k^1)} \mathbf{Y}(\mathbf{X}_k^1) \end{aligned} \tag{29}$$

$$\begin{aligned} t_{1k} &= \mathbf{W}_{1k}^{\mathrm{T}} \frac{(\hat{\mathbf{m}}_{1k} + \hat{\mathbf{m}}_2)}{2} = t - \frac{\mathbf{W}^{\mathrm{T}}(\mathbf{X}_k^1 - \hat{\mathbf{m}}_1)}{2(N_1 - 1)} - \frac{\mathbf{Y}^{\mathrm{T}}(\mathbf{X}_k^1)\, \hat{\mathbf{m}}}{2(N_1 - 1)\, \nu(\mathbf{X}_k^1)} + \frac{\beta(\mathbf{X}_k^1)}{2(N_1 - 1)^2\, \nu(\mathbf{X}_k^1)} \\ &\quad + \frac{\alpha Z(\mathbf{X}_k^1)}{\nu(\mathbf{X}_k^1)} \mathbf{Y}^{\mathrm{T}}(\mathbf{X}_k^1)\, \hat{\mathbf{m}} - \frac{\alpha Z(\mathbf{X}_k^1)\, \beta(\mathbf{X}_k^1)}{2(N_1 - 1)\, \nu(\mathbf{X}_k^1)} . \end{aligned} \tag{30}$$

Equations 29 and 30 can be used to compute $\mathbf{W}_{1k}$ and t_{1k} from $\mathbf{W}$ and t every time that a pattern $\mathbf{X}_k^1$ is left out from class ω_1 and the pattern $\mathbf{X}_k^1$ is tested. Similarly, recursive expressions can be derived when a pattern $\mathbf{X}_k^2$ is left out from class ω_2. It is to be noted that the matrix $\hat{\mathbf{S}}_{\mathbf{W}}$ needs to be computed and inverted twice, once when patterns from class ω_1 are left out and again when patterns from class ω_2 are left out.

Selection of an Optimal Threshold

This section considers the problem of finding the optimum threshold, t, to achieve minimum probability of error for the projected patterns onto Fisher's direction. The patterns in classes ω_i are assumed to be normally distributed, i.e., $p(\mathbf{X}|\omega_i) \sim N(\mathbf{m}_i, \boldsymbol{\Sigma}_i)$. Let y be the projection of pattern, $\mathbf{X}$, onto Fisher's direction, $\mathbf{W}$; i.e.,

$$y = \mathbf{W}^{\mathrm{T}}\mathbf{X}. \tag{31}$$

Because **X** is normally distributed, y is also normally distributed (for a constant **W**); i.e.,

$$p(y|\omega_i) \sim N(\mu_i, \sigma_i^2), i = 1, 2 \tag{32}$$

where

$$\mu_i = \mathbf{W}^T m_i \tag{33}$$

and

$$\sigma_i^2 = \mathbf{W}^T \mathbf{\Sigma}_i \mathbf{W}. \tag{34}$$

If the decision rule is used, decide $y \epsilon \omega_1$ if $y \geq t$; otherwise, decide $y \epsilon \omega_2$, the probability of error incurred, can be written as

$$P_e = P_1 \int_{-\infty}^{t} p(y|\omega_1)dy + P_2 \int_{t}^{\infty} p(y|\omega_2)dy = P_1 \int_{-\infty}^{\frac{t-\mu_1}{\sigma_1}} \phi(\zeta)d\zeta + P_2 \int_{\frac{t-\mu_2}{\sigma_2}}^{\infty} \phi(\zeta)d\zeta \tag{35}$$

where $\phi(\zeta) = 1/\sqrt{2\pi} \exp(-\frac{1}{2}\zeta^2)$ and P_i are the *a priori* probabilities of the classes ω_i, $i = 1, 2$. On differentiating Equation 35 with respect to t, the following is obtained:

$$\frac{\partial P_e}{\partial t} = P_1 \phi\left(\frac{t-\mu_1}{\sigma_1}\right)\frac{1}{\sigma_1} - P_2 \phi\left(\frac{t-\mu_2}{\sigma_2}\right)\frac{1}{\sigma_2}. \tag{36}$$

Equating $\partial P_e/\partial t$ to zero and then simplifying it, one obtains

$$\left(\frac{t-\mu_2}{\sigma_2}\right)^2 - \left(\frac{t-\mu_1}{\sigma_1}\right)^2 = 2 \log\left(\frac{P_2}{P_1}\frac{\sigma_1}{\sigma_2}\right). \tag{37}$$

The following cases are considered:

Case 1: $P_1 = P_2$, $\sigma_1 = \sigma_2$

From Equation 37, the optimum t that minimizes the probability of error when the patterns are projected onto Fisher's direction can be obtained as

$$t = \frac{\mu_1 + \mu_2}{2}. \tag{38}$$

It is seen that Equations 13 and 38 are identical.

Case 2: $P_1 \neq P_2$, $\sigma_1 = \sigma_2 = \sigma$

In this case, the optimum value of threshold t can be obtained from Equation 37 as

$$t = \frac{\sigma^2}{(\mu_1 - \mu_2)} \log\left(\frac{P_2}{P_1}\right) + \left(\frac{\mu_1 + \mu_2}{2}\right). \tag{39}$$

Case 3: $P_1 \neq P_2$, $\sigma_1 \neq \sigma_2$

On simplification, the following is obtained from Equation 37:

$$t^2 + \frac{(2\mu_1\sigma_2^2 - 2\mu_2\sigma_1^2)}{(\sigma_1^2 - \sigma_2^2)} t + \frac{(\sigma_1^2\mu_2^2 - \sigma_2^2\mu_1^2)}{(\sigma_1^2 - \sigma_2^2)} - \frac{2\sigma_1^2\sigma_2^2}{(\sigma_1^2 - \sigma_2^2)} \log\left(\frac{P_2}{P_1}\frac{\sigma_1}{\sigma_2}\right) = 0. \tag{40}$$

This is a quadratic equation of the form $at^2 + bt + c = 0$. The discriminant of the equation $\eta = b^2 - 4ac$ can be shown to be

$$\eta = \frac{4}{\left(\frac{\sigma_1}{\sigma_2} - \frac{\sigma_2}{\sigma_1}\right)^2}\left[(\mu_1 - \mu_2)^2 + 2(\sigma_1^2 - \sigma_2^2) \log\left(\frac{P_2}{P_1}\frac{\sigma_1}{\sigma_2}\right)\right]. \tag{41}$$

From Equation 41, it is seen that when $P_1 = P_2$, η is always positive, thus giving real roots for Equation 40. Even when $P_1 \neq P_2$, if η is positive, real roots are obtained for t. The η is negative when there exists no real threshold that minimizes the probability of error. Equation 40 gives two roots for t. To obtain t that minimizes P_e, differentiating Equation 36 with respect to t results in the following:

$$\frac{\partial^2 P_e}{\partial t^2} = \left(P_2 \frac{1}{\sigma_2^2} \left(\frac{t-\mu_2}{\sigma_2} \right) \phi \left(\frac{t-\mu_2}{\sigma_2} \right) - P_1 \frac{1}{\sigma_1^2} \left(\frac{t-\mu_1}{\sigma_1} \right) \phi \left(\frac{t-\mu_1}{\sigma_1} \right) \right]. \tag{42}$$

Let t_1 and t_2 be the roots of Equation 40, where

$$t_1 = (-b + \sqrt{\eta})/2,$$

$$t_2 = (-b - \sqrt{\eta})/2, \text{ and}$$

$$b = (2\,\mu_1\sigma_2^2 - 2\,\mu_2\sigma_1^2)/(\sigma_1^2 - \sigma_2^2).$$

It is shown in Appendix B that t_1 is the desired threshold which gives positive value for Equation 42. Using the results of the last section, one can update the threshold, t, for use with the leave-one-out method because it is a function of means and covariance matrices.

Generalization of the Fisher Classifier to Multiple Classes

Rewriting Equations 12 and 13 in terms of the discriminant functions $g_i(\mathbf{X}) = \mathbf{V}_i^{\mathrm{T}}\mathbf{X} + v_i$, $i = 1, 2$, the following decision rule is implemented:

$$\text{Decide } \mathbf{X}\epsilon\omega_1 \text{ if } g_1(\mathbf{X}) > g_2(\mathbf{X}) \tag{43}$$

$$\text{Decide } \mathbf{X}\epsilon\omega_2 \text{ if } g_1(\mathbf{X}) < g_2(\mathbf{X}). \tag{44}$$

Thus

$$\mathbf{V}_i = \hat{\mathbf{S}}_{\mathrm{W}}^{-1}\,\hat{\mathbf{m}}_i \tag{45}$$

and

$$v_i = -\hat{\mathbf{m}}_i^{\mathrm{T}}\,\hat{\mathbf{S}}_{\mathrm{W}}^{-1}\,\frac{(\hat{\mathbf{m}}_1 + \hat{\mathbf{m}}_2)}{2}\,. \tag{46}$$

It is seen that Equations 43 to 46 implement the decision rule of Equation 6. This suggests the definition of discriminant functions for an M-class problem as

$$g_i(\mathbf{X}) = \mathbf{V}_i^{\mathrm{T}}\mathbf{X} + v_i,\ i = 1, 2, \ldots, M$$

where

$$\mathbf{V}_i = \hat{\mathbf{S}}_{\mathrm{W}}^{-1}\,\hat{\mathbf{m}}_i$$

$$v_i = -\hat{\mathbf{m}}_i^{\mathrm{T}}\,\hat{\mathbf{S}}_{\mathrm{W}}^{-1}\frac{(\hat{\mathbf{m}}_1 + \hat{\mathbf{m}}_2 + \ldots + \hat{\mathbf{m}}_M)}{M}\,,\ i = 1, 2, \ldots, M \tag{47}$$

$$\hat{\mathbf{S}}_{\mathrm{W}} = \hat{\boldsymbol{\Sigma}}_1 + \hat{\boldsymbol{\Sigma}}_2 + \ldots + \hat{\boldsymbol{\Sigma}}_M.$$

Then the decision rule is the following: Decide $\mathbf{X}\epsilon\omega_i$ if

$$g_i(\mathbf{X}) > g_j(\mathbf{X}).\ j = 1, 2, \ldots, M \text{ and } j \neq i \tag{48}$$

Computational Expressions for the Leave-One-Out Method in a Multiclass Case

This section presents computational expressions for the leave-one-out method for updating $\mathbf{V}_i$ and v_i. Let there be M classes. Consider the case when a pattern $\mathbf{X}_k^1$ from class ω_1 is left out. Define the means and covariance matrices of the total pattern set as

$$\hat{\mathbf{m}}_i = \frac{1}{N_i}\sum_{j=1}^{N_i}\mathbf{X}_j^i,\ i = 1, 2, \ldots, M$$

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{1}{(N_1 - 2)}\sum_{j=1}^{N_1}(\mathbf{X}_j^1 - \hat{\mathbf{m}}_1)\,(\mathbf{X}_j^1 - \hat{\mathbf{m}}_1)^{\mathrm{T}} \tag{49}$$

$$\hat{\boldsymbol{\Sigma}}_i = \frac{1}{(N_1 - 1)}\sum_{j=1}^{N_i}(\mathbf{X}_j^i - \hat{\mathbf{m}}_i)\,(\mathbf{X}_j^i - \hat{\mathbf{m}}_i)^{\mathrm{T}},\ i = 2, \ldots, M.$$

Let $\hat{\mathbf{S}}_W = \hat{\boldsymbol{\Sigma}}_1 + \hat{\boldsymbol{\Sigma}}_2 + \ldots + \hat{\boldsymbol{\Sigma}}_M$. Compute $\mathbf{V}_i$ and v_i, $i = 1, 2, \ldots, M$ as

$$\begin{aligned} \mathbf{V}_i &= \hat{\mathbf{S}}_W^{-1}\, \hat{\mathbf{m}}_i \\ v_i &= -\hat{\mathbf{m}}_i^T\, \hat{\mathbf{S}}_W^{-1} \frac{(\hat{\mathbf{m}}_1 + \hat{\mathbf{m}}_2 + \ldots + \hat{\mathbf{m}}_M)}{M} \end{aligned} \tag{50}$$

When the pattern $\mathbf{X}_k^1$ from class ω_1 is left out, Fisher's parameters are computed as

$$\begin{aligned} \mathbf{V}_1(\mathbf{X}_k^1) &= \hat{\mathbf{S}}_{W1k}^{-1}\, \hat{\mathbf{m}}_{1k} \\ v_1(\mathbf{X}_k^1) &= -\hat{\mathbf{m}}_{1k}^T\, \hat{\mathbf{S}}_{W1k}^{-1} \frac{(\hat{\mathbf{m}}_{1k} + \hat{\mathbf{m}}_2 + \ldots + \hat{\mathbf{m}}_M)}{M} \\ \mathbf{V}_i(\mathbf{X}_k^1) &= \hat{\mathbf{S}}_{W1k}^{-1}\, \hat{\mathbf{m}}_i,\ i = 2, \ldots, M \\ v_i(\mathbf{X}_k^1) &= -\hat{\mathbf{m}}_i^T\, \hat{\mathbf{S}}_{W1k} \frac{(\hat{\mathbf{m}}_{1k} + \hat{\mathbf{m}}_2 + \ldots + \hat{\mathbf{m}}_M)}{M},\ i = 2, \ldots, M \end{aligned} \tag{51}$$

where

$$\begin{aligned} \hat{\mathbf{m}}_{1k} &= \frac{1}{(N_1 - 1)} \sum_{\substack{j=1 \\ \neq k}}^{N_1} \mathbf{X}_j^1 \\ \hat{\boldsymbol{\Sigma}}_{1k} &= \frac{1}{(N_1 - 2)} \sum_{\substack{j=1 \\ \neq k}}^{N_1} (\mathbf{X}_j^1 - \hat{\mathbf{m}}_1)(\mathbf{X}_j^1 - \hat{\mathbf{m}}_1)^T \\ \hat{\mathbf{S}}_{W1k} &= \hat{\boldsymbol{\Sigma}}_{1k} + \hat{\boldsymbol{\Sigma}}_2 + \ldots + \hat{\boldsymbol{\Sigma}}_M. \end{aligned} \tag{52}$$

The $\hat{\mathbf{m}}_i$ and $\hat{\boldsymbol{\Sigma}}_i$, $i = 2, \ldots, M$ are defined as in Equation 49. One obtains recursive relations for Fisher's parameters as follows:

$$\begin{aligned} \mathbf{V}_1(\mathbf{X}_k^1) &= \mathbf{V}_1 + \frac{\alpha d_1}{\nu(\mathbf{X}_k^1)} \mathbf{Y}(\mathbf{X}_k^1) - \frac{1}{(N_1 - 1)\, \nu(\mathbf{X}_k^1)} \mathbf{Y}(\mathbf{X}_k^1) \\ v_1(\mathbf{X}_k^1) &= v_1 - \frac{\alpha}{\nu(\mathbf{X}_k^1)} d_1 d_2 + \frac{d_2}{(N_1 - 1)\, \nu(\mathbf{X}_k^1)} + \frac{d_1}{2(N_1 - 1)} + \frac{\alpha\beta(\mathbf{X}_k^1)}{2(N_1 - 1)\, \nu(\mathbf{X}_k^1)} d_1 - \frac{1}{2(N_1 - 1)^2} \frac{\beta(\mathbf{X}_k^1)}{\nu(\mathbf{X}_k^1)} \end{aligned} \tag{53}$$

$$\begin{aligned} V_i(\mathbf{X}_k^1) &= \mathbf{V}_i + \frac{\alpha e_i}{\nu(\mathbf{X}_k^1)} \mathbf{Y}(\mathbf{X}_k^1),\ i = 2, \ldots, M \\ v_i(\mathbf{X}_k^1) &= v_i - \frac{\alpha e_i d_2}{\nu(\mathbf{X}_k^1)} + \frac{e_i}{2(N_1 - 1)\, \nu(\mathbf{X}_k^1)},\ i = 2, \ldots, M \end{aligned} \tag{54}$$

where

$$\begin{aligned} \alpha &= \frac{N_1}{(N_1 - 1)\,(N_1 - 2)} \\ \mathbf{Y}(\mathbf{X}_k^1) &= \hat{\mathbf{S}}_W^{-1}\, (\mathbf{X}_k^1 - \hat{\mathbf{m}}_1) \\ \boldsymbol{\beta}(\mathbf{X}_k^1) &= (\mathbf{X}_k^1 - \hat{\mathbf{m}}_1)^T\, \hat{\mathbf{S}}_W^{-1}\, (\mathbf{X}_k^1 - \hat{\mathbf{m}}_1) \\ \nu(\mathbf{X}_k^1) &= 1 - \alpha\beta(\mathbf{X}_k^1) \\ d_1 &= \mathbf{Y}^T(\mathbf{X}_k^1)\, \hat{\mathbf{m}}_1 \\ d_2 &= \mathbf{Y}^T(\mathbf{X}_k^1) \frac{(\hat{\mathbf{m}}_1 + \hat{\mathbf{m}}_2 + \ldots + \hat{\mathbf{m}}_M)}{M} = \mathbf{Y}^T(\mathbf{X}_k^1)\, \hat{\mathbf{m}} \\ e_i &= \mathbf{Y}^T(\mathbf{X}_k^1)\, \hat{\mathbf{m}}_i,\ i = 2, \ldots, M. \end{aligned}$$

Recursive relations can be obtained similarly when a pattern $\mathbf{X}_k^i$ from class ω_i is left out. It is to be noted that the matrix $\hat{\mathbf{S}}_W$ is to be inverted once for each class. The use of these recursive relations results in a computationally efficient way of implementing the leave-one-out method.

Conclusions

The Fisher classifier is one of the simplest and most widely used linear classifiers. Recently, considerable interest in its application for the classification of multispectral data acquired by Landsat has been expressed. Acquiring labels of the training patterns is expensive, and in many cases the probability of error is to be estimated in addition to the designing of a classifier. (For example, in remote sensing, a separate set of labeled patterns is used for estimating the probability of error.) Hence, in practical applications, it is advantageous to use the available labeled patterns more effectively.

This paper has presented computational expressions for estimating the probability of error using the leave-one-out method. Thus, the available labeled patterns can be used effectively both for designing the classifier and for estimating the probability of error. Because the classification accuracy depends on the threshold used with the Fisher classifier, expressions for optimal threshold for minimizing the probability of error in Fisher's direction are presented.

References

Bodewig, E., 1959. *Matrix Calculus*, Amsterdam, North Holland.

Chittineni, C. B., 1972. On the Fisher Criterion and Divergence in Pattern Recognition, *Internat. J. Control*, V.16, No. 1, pp. 205–207.

———, 1977. On the Estimation of Probability of Error, *Pattern Recognition*, V.9, No. 4, pp. 191–196.

Fisher, R. A., 1963. The Use of Multiple Measurements in Taxonomic Problems, *Ann. Eugenics*, V.7, Part II, pp. 179–188. (Also in *Contributions to Mathematical Statistics*, John Wiley and Sons, Inc., New York, 1963.)

Hills, M., 1966. Allocation Rules and Their Error Rates, *J. Royal Stat. Soc.*, Series B, V.28, pp. 1–31.

Lachenbruch, P. A., and M. R. Mickey, 1968. Estimation of Error Rates in Discriminant Analysis, *Technometrics*, V.10, pp. 1–11.

Misra, P. N., 1979. *Preliminary Results on Linear Classification to Identify Wheat and Estimation of Misclassification Errors*, IBM Report RES 23-65, International Business Machines Corporation (Houston, Texas).

Appendix A

Derivation of Matrix Relations

From Equation 14, one obtains

$$\hat{\mathbf{m}}_{1k} = \frac{1}{(N_1 - 1)}\left(\sum_{\substack{j=1 \\ \neq k}}^{N_1} \mathbf{X}_j^1\right) = \frac{1}{(N_1 - 1)}\left(\sum_{j=1}^{N_1} \mathbf{X}_j^1 - \mathbf{X}_k^1\right) = \frac{1}{(N_1 - 1)}(\mathbf{N}_1 \hat{\mathbf{m}}_1 - \mathbf{X}_k^1)$$

$$= \hat{\mathbf{m}}_1 - \frac{1}{(N_1 - 1)}(\mathbf{X}_k^1 - \hat{\mathbf{m}}_1) \tag{A-1}$$

thus obtaining Equation 18. From Equation 15,

$$\hat{\boldsymbol{\Sigma}}_{1k} = \frac{1}{N_1 - 2}\sum_{\substack{j=1 \\ \neq k}}^{N_1} (\mathbf{X}_j^1 - \hat{\mathbf{m}}_{1k})(\mathbf{X}_j^1 - \hat{\mathbf{m}}_{1k})^T$$

$$= \frac{1}{N_1 - 2}\left[\sum_{j=1}^{N_1} (\mathbf{X}_j^1 - \hat{\mathbf{m}}_{1k})(\mathbf{X}_j^1 - \hat{\mathbf{m}}_{1k})^T - (\mathbf{X}_k^1 - \hat{\mathbf{m}}_{1k})(\mathbf{X}_k^1 - \hat{\mathbf{m}}_{1k})^T\right] \tag{A-2}$$

Consider the following:

$$\sum_{j=1}^{N_1} (\mathbf{X}_j^1 - \hat{\mathbf{m}}_{1k})(\mathbf{X}_j^1 - \hat{\mathbf{m}}_{1k})^T = \sum_{j=1}^{N1}\left[\mathbf{X}_j^1 - \hat{\mathbf{m}}_1 + \frac{1}{N_1 - 1}(\mathbf{X}_k^1 - \hat{\mathbf{m}}_1)\right]\left[\mathbf{X}_j^1 - \hat{\mathbf{m}}_1 + \frac{1}{N_1 - 1}(\mathbf{X}_k^1 - \hat{\mathbf{m}}_1)\right]^T$$

$$= \sum_{j=1}^{N_1} (\mathbf{X}_j^1 - \hat{\mathbf{m}}_1)(\mathbf{X}_j^1 - \hat{\mathbf{m}}_1)^T + \frac{1}{(N_1 - 1)}(\mathbf{X}_k^1 - \hat{\mathbf{m}}_1)\sum_{j=1}^{N_1} (\mathbf{X}_j^1 - \hat{\mathbf{m}}_1)^T$$

$$+ \left[\sum_{j=1}^{N_1} (\mathbf{X}_j^1 - \hat{\mathbf{m}}_1)\right]\frac{1}{(N_1 - 1}(\mathbf{X}_k^1 - \hat{\mathbf{m}}_1)^T + \frac{N_1}{(N_1 - 1)^2}(\mathbf{X}_k^1 - \hat{\mathbf{m}}_1)(\mathbf{X}_k^1 - \hat{\mathbf{m}}_1)^T$$

$$= (N_1 - 2)\hat{\boldsymbol{\Sigma}}_1 + \frac{N_1}{(N_1 - 1)^2}(\mathbf{X}_k^1 - \hat{\mathbf{m}}_1)(\mathbf{X}_k^1 - \hat{\mathbf{m}}_1)^T \tag{A-3}$$

Consider

$$(\mathbf{X}_k^1 - \hat{\mathbf{m}}_{1k}) = \mathbf{X}_k^1 - \hat{\mathbf{m}}_1 + \frac{1}{(N_1 - 1)}(\mathbf{X}_k^1 - \hat{\mathbf{m}}_1) = \frac{N_1}{N_1 - 1}(\mathbf{X}_k^1 - \hat{\mathbf{m}}_1) \tag{A-4}$$

Substituting Equations A-3 and A-4 into A-2 results in the following:

$$\hat{\mathbf{\Sigma}}_{1k} = \frac{1}{(N_1 - 2)}\left[(N_1 - 2)\,\hat{\mathbf{\Sigma}}_1 + \frac{N_1}{(N_1 - 1)^2}(\mathbf{X}_k^1 - \hat{\mathbf{m}}_1)(\mathbf{X}_k^1 - \hat{\mathbf{m}}_1)^T - \frac{N_1^2}{(N_1 - 1)^2}(\mathbf{X}_k^1 - \hat{\mathbf{m}}_1)(\mathbf{X}_k^1 - \hat{\mathbf{m}}_1)^T\right]$$

$$= \hat{\mathbf{\Sigma}}_1 - \frac{N_1}{(N_1 - 1)(N_1 - 2)}\left[(\mathbf{X}_k^1 - \hat{\mathbf{m}}_1)(\mathbf{X}_k^1 - \hat{\mathbf{m}}_1)^T\right] \tag{A-5}$$

thus obtaining Equation 19.

Let $\mathbf{S} = \mathbf{\Sigma} - \alpha\mathbf{M}\mathbf{M}^T$, where $\mathbf{S}$ and $\mathbf{\Sigma}$ are nonsingular matrices and $\mathbf{M}$ is a vector. Then the inverse of $\mathbf{S}$ can be expressed in terms of the inverse of $\mathbf{\Sigma}$, as in Bodewig (1959):

$$\mathbf{S}^{-1} = \mathbf{\Sigma}^{-1} + \frac{\alpha\mathbf{\Sigma}^{-1}\mathbf{M}\mathbf{M}^T\mathbf{\Sigma}^{-1}}{1 - \alpha\mathbf{M}^T\mathbf{\Sigma}^{-1}\mathbf{M}} \tag{A-6}$$

thus obtaining Equation 22.

Appendix B

Derivation of the Optimal Threshold for the Case $P_1 \neq P_2$ and $\sigma_1 \neq \sigma_2$

Using Equations 40 and 41, the roots t_1 and t_2 of Equation 40 can be written as

$$t_1 = -b_1 + \eta_1\eta_2 \tag{B-1}$$

$$t_2 = -b_1 - \eta_1\eta_2 \tag{B-2}$$

where

$$b_1 = (\mu_1\sigma_2^2 - \mu_2\sigma_1^2)/(\sigma_1^2 - \sigma_2^2), \tag{B-3}$$

$$\eta_1 = (\sigma_1\sigma_2)/(\sigma_1^2 - \sigma_2^2), \text{ and} \tag{B-4}$$

$$\eta_2 = \left((\mu_1 - \mu_2)^2 + 2(\sigma_1^2 - \sigma_2^2)\log(P_2\sigma_1/P_1\sigma_2)\right)^{1/2}. \tag{B-5}$$

From Equations B-1, B-3, and B-4, we get the following:

$$\frac{(t_1 - \mu_2)}{\sigma_2} = \frac{\sigma_2(\mu_2 - \mu_1) + \sigma_1\eta_2}{(\sigma_1^2 - \sigma_2^2)} \tag{B-6}$$

$$\frac{(t_1 - \mu_1)}{\sigma_1} = \frac{\sigma_1(\mu_2 - \mu_1) + \sigma_2\eta_2}{(\sigma_1^2 - \sigma_2^2)}. \tag{B-7}$$

A sufficient condition that t_1 minimizes the probability of error P_e is

$$\frac{\partial^2 P_e}{\partial t^2}\Big|_{t=t_1} > 0. \tag{B-8}$$

Substituting Equations B-1, B-6, and B-7 in Equation 42 and using the condition B-8 yields

$$\frac{P_2}{\sigma_2^2}\frac{(\sigma_2(\mu_2 - \mu_1) + \sigma_1\eta_2)}{(\sigma_1^2 - \sigma_2^2)} > \frac{P_1}{\sigma_1^2}\frac{(\mu_2 - \mu_1)\sigma_1 + \sigma_2\eta_2}{(\sigma_1^2 - \sigma_2^2)}\exp\left(\log\left(\frac{P_2\sigma_1}{P_1\sigma_2}\right)\right) \tag{B-9}$$

On simplification, we get from B-9

$$\frac{1}{\sigma_1\sigma_2}\eta_2 > 0. \tag{B-10}$$

It is seen that the condition B-8 is satisifed when $\eta_2 > 0$. But there exists no real threshold when $\eta_2 < 0$. Thus, the optimal threshold that minimizes the probability of error is given by t_1 of Equation B-1. Proceeding similarly, it can easily be shown that the condition B-8 is not satisfied when $t = t_2$ and when $\eta_2 > 0$.

(Received 7 February 1981; revised and accepted 4 March 1982)