JSC-14870

Lockheed
Electronics
Company, Inc.

A SUBSIDIARY OF
LOCKHEED CORPORATION
1830 NASA Road 1, Houston, Texas 77058
Tel. 713-333-5411

NASA CR-
160263

TECHNICAL MEMORANDUM

EFFICIENT FEATURE SUBSET SELECTION WITH PROBABILISTIC DISTANCE CRITERIA

By

C. B. Chittineni

Approved By: T. C. Minter
T. C. Minter, Supervisor
Techniques Development
Section

May 1979

LEC-13355

CONTENTS

Section

Page

Appendix

TABLES

1. INTRODUCTION

Feature selection is one of the important problems in pattern recognition. Considerable interest has been shown on this problem in recent literature. Usually, the performance of the recognition system is expressed in terms of the probability of misrecognition P_e. Unfortunately, it is often difficult to obtain an analytical expression for P_e; and even if one can be obtained, it will usually be too complicated to permit analytical or numerical computation. Hence, certain probabilistic distance measures (ref. 1), which are easy to evaluate and manipulate, are used as criteria for the selection of effective features.

The distance measures that are normally used in practice are listed in table 1. Among these distance measures, divergence (refs. 2 to 5) and Bhattacharyya distance (refs. 6 and 7) are extensively investigated in the literature. These distance measures either provide bounds on the probability of error or give intuitive justification for the measure of separability between the classes. If the distributions of the patterns in the classes are assumed to be multivariate normal; i.e., if

$$p(X|\omega_i) \sim N(m_i, \Sigma_i),$$

closed-form expressions can be derived for the distance measures given in table 1. The closed-form expressions are listed in table 2.

For feature selection, the use of the distance measures is as follows. Suppose that r features are to be selected out of given S features. There are $\binom{S}{r}$ different combinations of r features. In a two-class case, for each feature subset one of the criteria given in table 2 is computed as a measure of effectiveness of the feature subset; and that feature subset is selected as the best, which extremizes the criterion. In a multiclass case (refs. 6 and 13), the distance measures are computed for the feature subset between all pairs of classes; and the maximum of either the minimum distance between class pairs or the mean value of the distance between class pairs is used as

TABLE 1.— PROBABILISTIC DISTANCE MEASURES

No.	Name	Reference	Definition of distance measure
1	Divergence	2 to 5	$J = \int [p(X\|\omega_1) - p(X\|\omega_2)] \log\left[\frac{p(X\|\omega_1)}{p(X\|\omega_2)}\right] dX$
2	Bhattacharyya distance	6, 7	$B = -\ell n \int [p(X\|\omega_1)p(X\|\omega_2)]^{\frac{1}{2}}dX$
3	Jeffreys-Matusita distance	8 to 10	$JM = \left[\int\left(\sqrt{p(X\|\omega_1)} - \sqrt{p(X\|\omega_2)}\right)^2 dX\right]^{\frac{1}{2}}$
4	Kullback-Leibler numbers	11	$KL12 = \int \ell n\left[\frac{p(X\|\omega_1)}{p(X\|\omega_2)}\right] p(X\|\omega_1)dX$
5	Mahalanobis distance	12	$\Delta = \left[(m_1 - m_2)^T \Sigma^{-1}(m_1 - m_2)\right]^{\frac{1}{2}}$

Notation:

$p(X|\omega_i)$ — Probability density functions of the patterns in classes ω_i

m_i — Mean of the patterns in class ω_i

$\int(\ )dX$ — Multivariate integral

Δ — For Mahalanobis distance, the distributions of the patterns in the classes ω_i are normal with means m_i and common variance matrix Σ.

TABLE 2.— DISTANCES BETWEEN TWO MULTIVARIATE NORMAL DENSITIES

No.	Name	Distance
1	Divergence	$J = \frac{1}{2}\mathrm{tr}\left(\Sigma_1^{-1}\Sigma_2 + \Sigma_2^{-1}\Sigma_1 - 2I\right) + \frac{1}{2}\mathrm{tr}\left[\left(\Sigma_1^{-1} + \Sigma_2^{-1}\right)(m_1 - m_2)(m_1 - m_2)^T\right]$
2	Bhattacharyya distance	$B = (1/8)(m_1 - m_2)^T \left(\frac{\Sigma_1 + \Sigma_2}{2}\right)^{-1} (m_1 - m_2) + \frac{1}{2}\ell n \left\{\frac{\det[\frac{1}{2}(\Sigma_1 + \Sigma_2)]}{[\det(\Sigma_1)\det(\Sigma_2)]^{\frac{1}{2}}}\right\}$
3	Jeffreys-Matusita distance	$JM = \left[2\left\{1 - \frac{[\det(\Sigma_1)\det(\Sigma_2)]^{\frac{1}{4}}}{[\det\{\frac{1}{2}(\Sigma_1 + \Sigma_2)\}]^{\frac{1}{2}}} \exp\left[-(1/8)(m_1 - m_2)\left(\frac{\Sigma_1 - \Sigma_2}{2}\right)^{-1}(m_1 - m_2)\right]\right\}\right]$
4	Kullback-Leibler number	$KL12 = \frac{1}{2}\ell n\left[\frac{\det(\Sigma_1)}{\det(\Sigma_2)}\right] + \frac{1}{2}\mathrm{tr}\left[\Sigma_1\left(\Sigma_2^{-1} - \Sigma_1^{-1}\right)\right] + \frac{1}{2}\mathrm{tr}\left[\Sigma_2^{-1}(m_1 - m_2)(m_1 - m_2)^T\right]$
5	Mahalanobis distance	$= \left[(m_1 - m_2)^T\Sigma^{-1}(m_1 - m_2)\right]^{\frac{1}{2}} (\Sigma_1 = \Sigma_2 = \Sigma)$

a measure of the feature subset's effectiveness. Because the complete criterion function is to be computed for each feature subset, it is computationally very inefficient.

The purpose of this paper is to derive recursive relations for the criteria listed in table 2. That is, expressions are derived for the change in the criteria when a feature is deleted from the current feature subset. Expressions are also derived for the change in the criteria when a feature is added to the current feature subset. A combinatorial algorithm (ref. 14) is presented; it generates all possible r feature combinations out of given S features with a single feature change at each step in Sc_r steps. This algorithm and the recursive relations provide an efficient method of choosing the best fe.cure subset out of all possible feature subsets. The paper is organized as follows.

In section 2, recursive relations are derived for computing the distance measures when a feature is added to the current feature subset. In section 3, recursive expressions are developed for the computation of distance measures when a feature is deleted from the current feature subset. In section 4, a combinatorial algorithm is presented for generating all possible r combinations of S in $\binom{S}{r}$ steps with a single change at each step. Matrix relations used in the paper are derived in the appendices.

2. RECURSIVE RELATIONS FOR DISTANCE MEASURES WHEN A FEATURE IS ADDED

In this section, expressions are developed for recursively updating the distance measures (presented in section 1) when a feature is added to the current feature subset. Let the current feature subset contain features $x_1, x_2, \cdots, x_{r-1}$. The pattern X_{r-1} containing these features is represented as

$$X_{r-1} = (x_1, x_2, \cdots, x_{r-1})^T \tag{1}$$

Let $m_{r-1,i}$ and $\Sigma_{r-1,i}$ be the means and covariance matrices of the patterns X_{r-1} in class i, i = 1, 2, ···, M, where M is the number of classes. Let a feature x_r be added to the current feature subset. Then let the mean and covariance matrix of the pattern X_r in class i be $m_{r,i}$ and $\Sigma_{r,i}$. The $m_{r-1,i}$ and $m_{r,i}$; $\Sigma_{r-1,i}$ and $\Sigma_{r,i}$ are related as follows.

$$m_{r,i} = \begin{bmatrix} m_{r-1,i} \\ \mu_{r,i} \end{bmatrix} \tag{2}$$

and

$$\Sigma_{r,i} = \left[\begin{array}{c:c} \Sigma_{r-1,i} & \phi_{r-1,i} \\ \hdashline \phi^T_{r-1,i} & \sigma_{r,i} \end{array}\right] \tag{3}$$

A careful examination of table 2 shows that all the criteria listed contain terms such as determinant of a covariance matrix, inverse of a covariance matrix, and trace of the product of two matrices. In appendix A, recursive expressions for these component terms are developed. In the following, these relations are used to develop expressions for the recursive computation of the distance measures listed in table 2 when a feature x_r is added to the current feature subset $x_1, x_2, \cdots, x_{r-1}$.

2.1 DIVERGENCE

From table 2, the divergence between two Gaussian distributed pattern classes can be written as

$$J_r = \frac{1}{2}\text{tr}\left(\Sigma^{-1}_{r,1}\Sigma_{r,2} + \Sigma^{-1}_{r,2}\Sigma_{r,1} - 2I\right) + \frac{1}{2}\text{tr}\left[\left(\Sigma^{-1}_{r,1} + \Sigma^{-1}_{r,2}\right)\left(m_{r,1} - m_{r,2}\right)\left(m_{r,1} - m_{r,2}\right)^T\right] \tag{4}$$

From equations (A-6) and (A-8), the following is obtained.

$$\operatorname{tr}\left(\Sigma^{-1}_{r,1}\Sigma_{r,2} + \Sigma^{-1}_{r,2}\Sigma_{r,1} - 2I\right) = \operatorname{tr}\left(\Sigma^{-1}_{r-1,1}\Sigma_{r-1,2} + \Sigma^{-1}_{r-1,2}\Sigma_{r-1,1} - 2I\right)$$

$$+ \delta_{r,2}\alpha_{1|2} + \delta_{r,1}\alpha_{2|1} - 2 \qquad (5)$$

Let

$$m_{r,12} = m_{r,1} - m_{r,2}$$

$$= \begin{bmatrix} m_{r-1,1} \\ \mu_{r,1} \end{bmatrix} - \begin{bmatrix} m_{r-1,2} \\ \mu_{r,2} \end{bmatrix} = \begin{bmatrix} m_{r-1,12} \\ \mu_{r,12} \end{bmatrix} \qquad (6)$$

From equations (6) and (A-1), the following equation is derived:

$$m^T_{r,12}\Sigma^{-1}_{r,1}m_{r,12} = m^T_{r-1,12}\Sigma^{-1}_{r-1,1}m_{r-1,12} + \delta_{r,1}(m^T_{r-1,12}\Theta_{r-1,1})^2$$

$$- \mu_{r,12}\delta_{r,1}(m^T_{r-1,12}\Theta_{r-1,1}) - \mu_{r,12}\delta_{r,1}(m^T_{r-1,12}\Theta_{r-1,1})$$

$$+ \mu^2_{r,12}\delta_{r,1}$$

$$= m^T_{r-1,12}\Sigma^{-1}_{r-1,1}m_{r-1,12} + \delta_{r,1}(m^T_{r-1,12}\theta_{r-1,1} - \mu_{r,12})^2 \qquad (7)$$

Similar to equation (7), equation (8) can be written as follows:

$$m^T_{r,12}\Sigma^{-1}_{r,2}m_{r,12} = m^T_{r-1,12}\Sigma^{-1}_{r-1,2}m_{r-1,12} + \delta_{r,2}(m^T_{r-1,12}\Theta_{r-1,2} - \mu_{r,12})^2 \quad (8)$$

Combining equations (4), (5), (7), and (8) results in

$$J_r = J_{r-1} + \frac{1}{2}(\delta_{r,2}\alpha_{1|2} + \delta_{r,1}\alpha_{2|1} - 2)$$

$$+ \frac{1}{2}\,\delta_{r,1}(m^T_{r-1,12}\Theta_{r-1,1} - \mu_{r,12})^2 + \frac{1}{2}\,\delta_{r,2}(m^T_{r-1,12}\Theta_{r-1,2} - \mu_{r,12})^2 \quad (9)$$

2.2 BHATTACHARYYA DISTANCE

The Bhattacharyya distance between two Gaussian distributed pattern classes is given by

$$B_r = \frac{1}{4}(m_{r,1} - m_{r,2})^T(\Sigma_{r,1} + \Sigma_{r,2})^{-1}(m_{r,1} - m_{r,2})$$
$$+ \frac{1}{2}\ \ell n\left[\frac{\det\left[\frac{1}{2}(\Sigma_{r,1} + \Sigma_{r,2})\right]}{\{\det(\Sigma_{r,1})\det(\Sigma_{r,2})\}^{\frac{1}{2}}}\right] \tag{10}$$

Let $\Sigma_{r,1+2} = \Sigma_{r,1} + \Sigma_{r,2}$. Equation (11) can be written similarly to equation (7):

$$m^T_{r,12}\Sigma^{-1}_{r,1+2}m_{r,12} = m^T_{r-1,12}\Sigma^{-1}_{r-1,1+2}m_{r-1,12}$$
$$+ \delta_{r,1+2}(m^T_{r-1,12}\Theta_{r-1,1+2} - \mu_{r,12})^2 \tag{11}$$

From equation (A-5), the following equation is obtained:

$$\frac{1}{2}\ \ell n\left[\frac{\det\ \frac{1}{2}(\Sigma_{r,1} + \Sigma_{r,2})}{\{\det(\Sigma_{r,1})\det(\Sigma_{r,2})\}^{\frac{1}{2}}}\right] = \frac{1}{2}\ \ell n\left[\frac{\frac{1}{2^r}\det(\Sigma_{r,1+2})}{\{\det(\Sigma_{r,1})\det(\Sigma_{r,2})\}^{\frac{1}{2}}}\right]$$
$$= \frac{1}{2}\ \ell n\left[\frac{\det\left\{\frac{1}{2}(\Sigma_{r-1,1+2})\right\}}{\{\det(\Sigma_{r-1,1})\det(\Sigma_{r-1,2})\}^{\frac{1}{2}}}\right] + \frac{1}{2}\ \ell n\left[\frac{\{\delta_{r,1}\delta_{r,2}\}^{\frac{1}{2}}}{2\delta_{r,1+2}}\right] \tag{12}$$

From equations (10), (11), and (12), the recursive relation for the Bhattacharyya distance is obtained as

$$B_r = B_{r-1} + \frac{1}{4}\ \delta_{r,1+2}(m^T_{r-1,12}\Theta_{r-1,1+2} - \mu_{r,12})^2 + \frac{1}{2}\ \ell n\left(\frac{\{\delta_{r,1}\delta_{r,2}\}^{\frac{1}{2}}}{2\delta_{r,1+2}}\right) \tag{13}$$

2.3 JEFFREYS-MATUSITA DISTANCE

The Jeffreys-Matusita distance between two Gaussian distributed pattern classes is obtained from table 2:

$$JM_r = 2\left[1 - \left\{\frac{\left[\det(\Sigma_{r,1})\det(\Sigma_{r,2})\right]^{\frac{1}{4}}}{\left\{\det[\frac{1}{2}(\Sigma_{r,1} + \Sigma_{r,2})]\right\}^{\frac{1}{2}}} \exp[-\tfrac{1}{4}(m_{r,1} - m_{r,2})^T(\Sigma_{r,1} - \Sigma_{r,2})^{-1}(m_{r,1} - m_{r,2})]\right\}\right] \quad (14)$$

Let $\Sigma_{r,12} = \Sigma_{r,1} - \Sigma_{r,2}$.

Similar to equation (7), the following equation is derived:

$$m^T_{r,12}\Sigma^{-1}_{r,12}m_{r,12} = m^T_{r-1,12}\Sigma^{-1}_{r-1,12}m_{r-1,12} + \delta_{r,12}(m^T_{r-1,12}\Theta_{r-1,12} - \mu_{r,12})^2 \quad (16)$$

Let

$$C = \frac{(2\delta_{r,1+2})^{\frac{1}{2}}}{(\delta_{r,1}\delta_{r,2})^{\frac{1}{4}}} \exp\left[-\frac{1}{4}\,\delta_{r,12}(m^T_{r-1,12}\Theta_{r-1,12} - \mu_{r,12})^2\right] \quad (17)$$

From equations (A-5), (14), (16), and (17), the recursive relation for the Jeffreys-Matusita distance is obtained as

$$JM_r = C\ JM_{r-1} + 2(1 - C) \quad (18)$$

2.4 KULLBACK-LEIBLER NUMBERS

The Kullback-Leibler number between two Gaussian distributed pattern classes is given by

$$KL12_r = \frac{1}{2}\,\ell n\left[\frac{\det(\Sigma_{r,1})}{\det(\Sigma_{r,2}}\right] + \frac{1}{2}\,tr(\Sigma_{r,1}\Sigma^{-1}_{r,2} - I)$$

$$+ \frac{1}{2}(m_{r,1} - m_{r,2})^T\Sigma^{-1}_{r,2}(m_{r,1} - m_{r,2}) \quad (19)$$

From equations (7), (19), (A-5), and (A-6), the recursive relation for the Kullback-Leibler number is obtained as

$$KL12_r = KL12_{r-1} + \frac{1}{2} \ln\left(\frac{\delta_{r,2}}{\delta_{r,1}}\right) + \frac{1}{2}(\delta_{r,2}\alpha_{1|2} - 1)$$

$$+ \frac{1}{2} \delta_{r,2}(m^T_{r-1,12}\Theta_{r-1,2} - \mu_{r,2})^2 \qquad (20)$$

2.5 MAHALANOBIS DISTANCE

In Mahalanobis distance, Σ is usually taken as an average of the covariance matrices of the two pattern classes. Then it is defined as

$$\Delta^2_r = (m_{r,1} - m_{r,2})^T\left(\frac{\Sigma_{r,1} + \Sigma_{r,2}}{2}\right)^{-1} (m_{r,1} - m_{r,2}) \qquad (21)$$

From equations (7) and (21), the recursive expression for the Mahalanobis distance is obtained as

$$\Delta^2_r = \Delta^2_{r-1} + 2\delta_{r,1+2}(m^T_{r-1,12}\Theta_{r-1,1+2} - \mu_{r,12})^2 \qquad (22)$$

3. RECURSIVE RELATIONS FOR DISTANCE MEASURES WHEN A FEATURE IS DELETED

Recursive expressions are presented in this section for updating the distance measures given in section 1 when a feature is deleted from the current feature subset. Let the current feature subset contain features x_1, x_2, $\cdots$, x_r. The pattern X_r containing these features is represented as

$$X_r = (x_1, x_2, \cdots, x_r)^T$$

Let $m_{r,i}$ and $\Sigma_{r,i}$ be the mean and the covariance matrix of the pattern X_r in class i. Let the feature x_r be deleted from the current feature subset. Then let the mean and the covariance matrix of the patterns X_{r-1} in class i be $m_{r-1,i}$ and $\Sigma_{r-1,i}$. The $m_{r,i}$ and the $m_{r-1,i}$; $\Sigma_{r,i}$ and $\Sigma_{r-1,i}$ are related as in equations (2) and (3). Appendix B presents the derivations of the recursive relations for the determinant of a covariance matrix, the inverse of a covariance matrix, and the trace of the product of two matrices when a feature is deleted from the current feature subset. These relations are used in the following subsections in deriving expressions for recursively computing the distance measures when a feature is deleted from the current feature subset (section 1).

3.1 DIVERGENCE

The divergence between two Gaussian distributed pattern classes is given by equation (4). From equation (B-13), the following is obtained:

$$\mathrm{tr}(\Sigma_{r,1}\Sigma^{-1}_{r,2}) = \mathrm{tr}(\Sigma_{r-1,1}\Sigma^{-1}_{r-1,2}) + \frac{\xi'^{T}_{r,2}\Sigma_{r,1}\xi'_{r,2}}{\delta_{r,2}} \tag{23}$$

From equations (6), (B-1), and (B-8), the following is obtained:

$$\begin{aligned} m^{T}_{r,12}\Sigma^{-1}_{r,1}m_{r,12} &= m^{T}_{r-1,12}\psi_{r-1,1}m_{r-1,12} + \mu_{r,12}\xi^{T}_{r-1,1}m_{r-1,12} \\ &\quad + m^{T}_{r-1,12}\xi_{r-1,i}\mu_{r,12} + \mu^{2}_{r,12}\delta_{r,1} \\ &= m^{T}_{r-1,12}\Sigma^{-1}_{r-1,1}m_{r-1,12} + \frac{(m^{T}_{r-1,12}\xi_{r-1,1})^{2}}{\delta_{r,1}} \\ &\quad + 2\mu_{r,12}\xi^{T}_{r-1,1}m_{r-1,12} + \mu^{2}_{r,12}\delta_{r,1} \\ &= m^{T}_{r-1,12}\Sigma^{-1}_{r-1,1}m_{r-1,12} + \frac{1}{\delta_{r,1}}\left\{\begin{bmatrix} m^{T}_{r-1,12} & \mu_{r,12}\end{bmatrix}\begin{bmatrix}\xi_{r-1,1} \\ \delta_{r,1}\end{bmatrix}\right\}^{2} \\ &= m^{T}_{r-1,12}\Sigma^{-1}_{r-1,1}m_{r-1,12} + \frac{\left(m^{T}_{r,12}\xi'_{r,1}\right)^{2}}{\delta_{r,1}} \end{aligned} \tag{24}$$

Similar to equations (23) and (24), the following is obtained:

$$\mathrm{tr}\left(\Sigma_{r,2}\Sigma^{-1}_{r,1}\right) = \mathrm{tr}\left(\Sigma_{r-1,2}\Sigma^{-1}_{r-1,1}\right) + \frac{\xi'^{T}_{r,1}\Sigma_{r,2}\xi'_{r,1}}{\delta_{r,1}} \tag{25}$$

$$m^{T}_{r,12}\Sigma^{-1}_{r,2}m_{r,12} = m^{T}_{r-1,12}\Sigma^{-1}_{r-1,2}m_{r-1,12} + \frac{\left(m^{T}_{r,12}\xi'_{r,2}\right)^{2}}{\delta_{r,2}} \tag{26}$$

From equation (4) and equations (23) to (26), the following recursive relation is obtained for the computation of divergence.

$$J_{r-1} = J_r - \frac{\xi'^T_{r,2}\left(\Sigma_{r-1} + m_{r,12}m^T_{r,12}\right)\xi'_{r,2}}{2\delta_{r,2}} - \frac{\xi'^T_{r,1}\left(\Sigma_{r,2} + m_{r,12}m^T_{r,12}\right)\xi'_{r,1}}{2\delta_{r,1}} + 1 \tag{27}$$

3.2 BHATTACHARYYA DISTANCE

The Bhattacharyya distance between two Gaussian distributed pattern classes is given by equation (10). Similar to equation (24), equation (28) is obtained:

$$m^T_{r,12}\Sigma^{-1}_{r,1+2}m_{r,12} = m^T_{r-1,12}\Sigma^{-1}_{r-1,1+2}m_{r-1,12} + \frac{\left(m^T_{r,12}\xi'_{r,1+2}\right)^2}{\delta_{r,1+2}} \tag{28}$$

From equations (10), (28), and (B-9), a recursive expression for the computation of Bhattacharyya distance can be obtained as

$$B_{r-1} = B_r - \frac{1}{4}\,\frac{\left(m^T_{r,12}\xi'_{r,1+2}\right)^2}{\delta_{r,1+2}} - \frac{1}{2}\,\ln\left[\frac{\left(\delta_{r,1}\delta_{r,2}\right)^{\frac{1}{2}}}{2\delta_{r,1+2}}\right] \tag{29}$$

3.3 JEFFREYS-MATUSITA DISTANCE

The Jeffreys-Matusita distance between two Gaussian distributed pattern classes is given by equation (14). Equation (15) is used similarly to equation (24) to obtain

$$m^T_{r,12}\Sigma^{-1}_{r,12}m_{r,12} = m^T_{r-1,12}\Sigma^{-1}_{r-1,12}m_{r-1,12} + \frac{\left(m^T_{r,12}\xi'_{r,12}\right)^2}{\delta_{r,12}} \tag{30}$$

Let

$$C = \frac{\left(2\delta_{r,12}\right)^{\frac{1}{2}}}{\left(\delta_{r,1}\delta_{r,2}\right)^{\frac{1}{4}}}\exp\left[-\frac{1}{4}\,\frac{\left(m^T_{r,12}\xi'_{r,12}\right)^2}{\delta_{r,12}}\right] \tag{31}$$

From equations (15), (B-9), (30), and (31), a recursive expression for the computation of Jeffreys-Matusita distance when a feature is deleted from the current feature subset can be written as

$$JM_{r-1} = 2 + \frac{1}{C}[JM_r - 2] \tag{32}$$

3.4 KULLBACK-LEIBLER NUMBERS

From table 2, the Kullback-Leibler numbers between two Gaussian distributed pattern classes is given in equation (19). Equations (B-9), (23), and (26) in (19) can be used to write a recursive expression for the computation of the Kullback-Leibler number as follows:

$$KL12_{r-1} = KL12_r - \frac{1}{2} \frac{\xi'^T_{r,2}\left(\Sigma_{r,1} + m_{r,12}m^T_{r,12}\right)\xi'_{r,2}}{\delta_{r,2}} + \frac{1}{2} - \frac{1}{2} \ln\left(\frac{\delta_{r,2}}{\delta_{r,1}}\right) \tag{33}$$

3.5 MAHALANOBIS DISTANCE

The Mahalanobis distance, taking the covariance matrix in it as the average of the covariance matrices of the two pattern classes, can be written as

$$\Delta^2_r = 2(m_{r,1} - m_{r,2})^T(\Sigma_{r,1} + \Sigma_{r,2})^{-1}(m_{r,1} - m_{r,2}) \tag{34}$$

From equations (28) and (34), a recursive relation for the computation of Mahalanobis distance when a feature is deleted from the current feature subset is obtained:

$$\Delta^2_{r-1} = \Delta^2_r - \frac{2\left(m^T_{r,12}\xi'_{r,1+2}\right)^2}{\delta_{r,1+2}} \tag{35}$$

4. A COMBINATORIAL ALGORITHM FOR GENERATING ALL POSSIBLE COMBINATIONS

This section describes an algorithm for generating all possible r combinations out of S in Sc_r steps. At each step, a single change is made; i.e., one feature is deleted and one is added. The recursive relations developed in sections 2 and 3, coupled with this algorithm, can be effectively used to search for a best feature subset of r features out of all possible $\binom{S}{r}$ feature subsets using probabilistic distance measures as the criteria.

The initial combination may be any combination in which all the r-selected features are numbered consecutively. In the binary representation it means that all the r 1's are in one run in a vector of length S. For example, if $r = 3$ and $S = 5$, one may start with 11100 or 00111. The binary vector is denoted by A, and its *ith* component is A(i). Initially, all the components of A, except those of the last run, are marked. For example, if A = 00111000 (for $r = 3$ and $S = 8$), then it is marked as $\overline{00111}000$.

If a is a symbol, a^m stands for $\underbrace{aa \cdots a}$ m times. Let i be the highest index j such that A(j) is marked. A vector T(1), T(2), ···, T(S) of integers that satisfy the condition $|T(j)| \leq j$ for $j = 1, 2, \cdots, S$ is defined. Initially, $T(1) = 0$. If the initial combination is $(\overline{0})^p(\overline{1})^r 0^{S-r-p}$ where $S > r + p$, then $T(p + r) = -1$ and all the rest are immaterial. If the initial combination is $(\overline{0})^{S-r}1^r$, then $T(S - r) = -1$ and all the rest are immaterial. The changes T must undergo in each combination generation are described by subroutines α and β as follows.

α: (i) If $T(k) = 0$, then output A and halt.

(ii) If $T(k) > 0$, then $i \leftarrow T(k)$, output A and go to step 1 of the algorithm.

(iii) $i \leftarrow k - 1$. If $T(k) > -(k - 1)$, then $T(k - 1) \leftarrow T(k)$.

(iv) Output A and go to step 1 of the algorithm.

β: (i) $T(i) \leftarrow -(k + 1)$. If $T(k) \geq 0$, then $T(k + 1) \leftarrow T(k)$, output A, and go to step 1 of the algorithm.

(ii) $T(k + 1) \leftarrow k - 1$. If $T(k) > -(k - 1)$, then $T(k - 1) \leftarrow T(k)$.

(iii) Output A and go to step 1 of the algorithm.

Now the vector F(0), F(1), ···, F(S) is introduced as follows. If $A(m) = 1$ and if it is the rightmost element in a run of 1's, then F(m) is the index of the first 1 of this run. If not, F(m) is immaterial. Let ℓ be the index of the rightmost 1; that is, $\ell = \max\limits_{A(m) = 1} m.$

Now an algorithm for generating all possible combinations with a single change at each step can be described as follows. The initial conditions of the algorithm are illustrated as follows. Let $r = 3$, $S = 8$ with an initial $A = 01110000$. Then $i = 4$.

```
1.  $k \leftarrow i$. If $A(i) = 1$, go to step 8.
2.  $j \leftarrow F(\ell)$.
3.  $A(i) \leftarrow 1$, $A(j) \leftarrow 0$, and $F(k) \leftarrow k$. If $A(k - 1) = 1$ and $k > 1$, then $F(k) \leftarrow F(k - 1)$. $F(\ell) \leftarrow j + 1$, if $j < \ell$, go to step 5.
4.  $\ell \leftarrow i$. Perform $\alpha$.
5.  If $\ell < S$, go to step 7.
6.  $i \leftarrow j$. Perform $\beta$.
7.  $i \leftarrow \ell$. Perform $\beta$.
8.  $F(i - 1) \leftarrow F(i)$. If $\ell > i$, go to step 12.
9.  $A(i) \leftarrow 0$, $A(S) \leftarrow 1$, $F(S) \leftarrow S$, $\ell \leftarrow S$. If $i < S - 1$, go to step 11.
10. Perform $\alpha$.
11. $i \leftarrow S - 1$. Perform $\beta$.
12. $j \leftarrow F(\ell)$.
13. $A(i) \leftarrow 0$, $A(j - 1) \leftarrow 1$. $F(\ell) \leftarrow j - 1$. If $\ell < S$, go to step 17.
14. If $\ell + 1 \leftarrow j - 1$, go to step 16.
15. Perform $\alpha$.
16. $i \leftarrow j - 2$. Perform $\beta$.
17. $i \leftarrow \ell$. Perform $\beta$.
```

5. CONCLUSIONS

This paper considered probabilistic distance measures as criteria for feature subset evaluation. The measures discussed are divergence, Bhattacharyya distance, Jeffreys-Matusita distance, Kullback-Leibler numbers, and Mahalanobis distance.

The problem of finding the best feature subset is that of evaluating all possible feature subsets and selecting the one that extremizes the criteria. Recursive expressions are derived for computing the criteria as a change in the distance measures, both when a feature is added to the current feature subset and when a feature is deleted from the current feature subset. A combinatorial algorithm is presented for generating all possible r feature combinations from a given set of S features in $\binom{S}{r}$ steps with a change of a single feature at each step. These recursive expressions and the combinatorial algorithm provide an efficient way of finding by exhaustive search the best feature subset using the probabilistic distance measures as criteria. These expressions can also be used for finding the suboptimal feature subset using forward or backward sequential feature selection methods.

APPENDIX A

RECURSIVE MATRIX RELATIONSHIPS WHEN A FEATURE IS ADDED

In this appendix, recursive relations are derived for the inverse of a matrix, determinant of a matrix, and trace of the product of two matrices when a feature x_r is added to the current feature subset $x_1, x_2, \cdots, x_{r-1}$.

A.1 INVERSE OF A COVARIANCE MATRIX

It can be shown that the inverse of equation (3) can be written (ref. 15) as

$$\Sigma_{r,i}^{-1} = \begin{bmatrix} \Sigma_{r-1,i}^{-1} + \delta_{r,i}\Theta_{r-1,i}\Theta_{r-1,i}^{T} & -\delta_{r,i}\Theta_{r-1,i} \\ -\delta_{r,i}\Theta_{r-1,i}^{T} & \delta_{r,i} \end{bmatrix} \tag{A-1}$$

where

$$\frac{1}{\delta_{r,i}} = \sigma_{r,i} - \phi_{r-1,i}^{T}\Sigma_{r-1,i}^{-1}\phi_{r-1,i}$$

and

$$\Theta_{r-1,i} = \Sigma_{r-1,i}^{-1}\phi_{r-1,i} \tag{A-2}$$

A.2 DETERMINANT OF A COVARIANCE MATRIX

Let the matrix $\Sigma_{r,i}$ be partitioned as in equation (3). Consider a matrix B.

$$B = \begin{bmatrix} I & -\Sigma_{r-1,i}^{-1}\phi_{r-1,i} \\ 0 & I \end{bmatrix} \tag{A-3}$$

The determinant of matrix B is unity. Form a matrix $B^T\Sigma_{r,i}B$,

$$B^T\Sigma_{r,i}B = \begin{bmatrix} I & 0 \\ -\phi^T_{r-1,i}\Sigma^{-1}_{r-1,i} & I \end{bmatrix} \begin{bmatrix} \Sigma_{r-1,i} & \phi_{r-1,i} \\ \phi^T_{r-1,i} & \sigma_{r,i} \end{bmatrix} \begin{bmatrix} I & -\Sigma^{-1}_{r-1,i}\phi_{r-1,i} \\ 0 & I \end{bmatrix}$$

$$= \begin{bmatrix} \Sigma_{r-1,i} & 0 \\ 0 & \left(\sigma_{r,i} - \phi^T_{r-1,i}\Sigma^{-1}_{r-1,i}\phi_{r-1,i}\right)I \end{bmatrix} \qquad \text{(A-4)}$$

Taking the determinants on both sides of equation (A-4), one obtains the following:

$$\det(\Sigma_{r,i}) = (\sigma_{r,i} - \phi^T_{r-1,i}\Sigma^{-1}_{r-1,i}\phi_{r-1,i})\det(\Sigma_{r-1,i})$$

$$= \frac{1}{\delta_{r,i}} \det(\Sigma_{r-1,\ i}) \qquad \text{(A-5)}$$

A.3 TRACE OF THE PRODUCT OF TWO MATRICES

From equations (3) and (A-1), the following is obtained:

$$\begin{aligned} tr\left(\Sigma_{r,1}\Sigma^{-1}_{r,2}\right) &= tr\Big\{\Sigma_{r-1,1}\Sigma^{-1}_{r-1,2} + \delta_{r,2}\Sigma_{r-1,1}\Theta_{r-1,2}\Theta^T_{r-1,2} \\ &\quad - \delta_{r,2}\phi_{r-1,1}\Theta^T_{r-1,2} - \delta_{r,2}\phi^T_{r-1,1}\Theta_{r-1,2} + \sigma_{r,1}\delta_{r,2}\Big\} \\ &= tr\left[\Sigma_{r-1,1}\Sigma^{-1}_{r-1,2}\right] \\ &\quad + \delta_{r,2}\left[\Theta^T_{r-1,2}\Sigma_{r-1,1}\Theta_{r-1,2} - 2\phi^T_{r-1,1}\Theta_{r-1,2} + \sigma_{r,1}\right] \\ &= tr\left(\Sigma_{r-1,1}\Sigma^{-1}_{r-1,2}\right) + \delta_{r,2}\alpha_{1|2} \end{aligned} \qquad \text{(A-6)}$$

where

$$\alpha_{1|2} = \Theta^T_{r-1,2}\Sigma_{r-1,1}\Theta_{r-1,2} - 2\phi^T_{r-1,1}\Theta_{r-1,2} + \sigma_{r,1} \qquad \text{(A-7)}$$

The following equation is obtained similarly:

$$\mathrm{tr}\left(\Sigma_{r,2}\Sigma_{r,1}^{-1}\right) = \mathrm{tr}\left(\Sigma_{r-1,2}\Sigma_{r-1,1}^{-1}\right) + \delta_{r,1}\alpha_{2|1} \qquad \text{(A-8)}$$

where

$$\left.\begin{aligned} \frac{1}{\delta_{r,1}} &= \sigma_{r,1} - \phi_{r-1,1}^{T}\Sigma_{r-1,1}^{-1}\phi_{r-1,1} \\ \Theta_{r-1,1} &= \Sigma_{r-1,1}^{-1}\phi_{r-1,1} \\ \alpha_{2|1} &= \Theta_{r-1,1}^{T}\Sigma_{r-1,2}\Theta_{r-1,1} - 2\phi_{r-1,2}^{T}\Theta_{r-1,1} + \sigma_{r,2} \end{aligned}\right\} \qquad \text{(A-9)}$$

APPENDIX B

RECURSIVE MATRIX RELATIONSHIPS WHEN A FEATURE IS DELETED

This appendix derives recursive expressions when a feature x_r is deleted from the current feature subset $x_1, x_2, \cdots, x_r$ for the inverse of a covariance matrix, determinant of a covariance matrix, and trace of the product of two matrices.

B.1 INVERSE OF A COVARIANCE MATRIX

Let the inverse of the covariance matrix $\Sigma_{r,i}$ of equation (3) be represented by

$$\Sigma_{r,i}^{-1} = \begin{bmatrix} \psi_{r-1,i} & \xi_{r-1,i} \\ \xi_{r-1,i}^{T} & \delta_{r,i} \end{bmatrix} \tag{B-1}$$

Since $\Sigma_{r,i}^{-1}$ is the inverse of $\Sigma_{r,i}$, one has

$$\Sigma_{r,i}^{-1}\Sigma_{r,i} = I \tag{B-2}$$

From equations (3), (B-1), and (B-2), the following are obtained:

$$\psi_{r-1,i}\Sigma_{r-1,i} + \xi_{r-1,i}\phi_{r-1,i}^{T} = I \tag{B-3}$$

$$\psi_{r-1,i}\phi_{r-1,i} + \sigma_{r,i}\xi_{r-1,i} = 0 \tag{B-4}$$

$$\xi_{r-1,i}^{T}\Sigma_{r-1,i} + \delta_{r,i}\phi_{r-1,i}^{T} = 0 \tag{B-5}$$

$$\xi_{r-1,i}^{T}\phi_{r-1,i} + \delta_{r,i}\sigma_{r,i} = 1 \tag{B-6}$$

noting that $\xi_{r-1,i}\phi_{r-1,i}^{T}$ is a matrix of unit rank, one gets from equation (B-3) (ref. 15),

$$\Sigma_{r-1,i}^{-1} = \psi_{r-1,i} + \frac{\xi_{r-1,i}\phi_{r-1,i}^{T}\psi_{r-1,i}}{1 - \phi_{r-1,i}^{T}\xi_{r-1,i}} \tag{B-7}$$

Using equations (B-4) and (B-6) in (B-7) yields

$$\Sigma^{-1}_{r-1,i} = \psi_{r-1,i} - \frac{\xi_{r-1,i}\xi^{T}_{r-1,i}}{\delta_{r,i}} \tag{B-8}$$

B.2 DETERMINANT OF A COVARIANCE MATRIX

From equation (A-5), $\det(\Sigma_{r,1})$ and $\det(\Sigma_{r-1,i})$ are related as

$$\det(\Sigma_{r-1,i}) = \delta_{r,i}\det(\Sigma_{r,i}) \tag{B-9}$$

B.3 TRACE OF THE PRODUCT OF TWO MATRICES

From equations (3), (B-1), and B-8), one obtains

$$\begin{aligned} \mathrm{tr}\left(\Sigma_{r,1}\Sigma^{-1}_{r,2}\right) &= \mathrm{tr}\left[\Sigma_{r-1,1}\psi_{r-1,2} + \phi_{r-1,1}\xi^{T}_{r-1,2} + \phi^{T}_{r-1,1}\xi_{r-1,2} + \sigma_{r,1}\delta_{r,2}\right] \\ &= \mathrm{tr}\left(\Sigma_{r-1,1}\Sigma^{-1}_{r-1,2}\right) + \mathrm{tr}\left(\frac{\Sigma_{r-1,1}\xi_{r-1,2}\xi^{T}_{r-1,2}}{\delta_{r,2}}\right) \\ &\quad + 2\phi^{T}_{r-1,1}\xi_{r-1,2} + \sigma_{r,1}\delta_{r,2} \end{aligned} \tag{B-10}$$

Let

$$\xi'_{r,2} = \begin{bmatrix} \xi_{r-1,2} \\ \delta_{r,2} \end{bmatrix} \tag{B-11}$$

Consider

$$\begin{aligned} \frac{\xi'^{T}_{r,2}\Sigma_{r,1}\xi'_{r,2}}{\delta_{r,2}} &= \mathrm{tr}\left[\frac{\Sigma_{r,1}\xi'_{r,2}\xi'^{T}_{r,2}}{\delta_{r,2}}\right] \\ &= \mathrm{tr}\left[\frac{\Sigma_{r-1,1}\xi_{r-1,2}\xi^{T}_{r-1,2}}{\delta_{r,2}}\right] + 2\phi^{T}_{r-1,1}\xi_{r-1,2} + \sigma_{r,1}\delta_{r,2} \end{aligned} \tag{B-12}$$

From equations (B-10) and (B-12), one obtains the required recursive expression,

$$\mathrm{tr}\left(\Sigma_{r-1,1}\Sigma_{r-1,2}^{-1}\right) = \mathrm{tr}\left(\Sigma_{r,1}\Sigma_{r,2}^{-1}\right) - \frac{\xi_{r,2}^{\prime T}\Sigma_{r,1}\xi_{r,2}^{\prime}}{\delta_{r,2}} \qquad \text{(B-13)}$$

APPENDIX C

REFERENCES

1. Wacker, A. G.; and Landgrebe, D. A.: The Minimum Distance Approach to Classification. LARS information note 100771, Oct. 1971.

2. Kailath, T.: The Divergence and Bhattacharyya Distance Measures in Signal Selection. IEEE Trans. Comm. Tech., vol. COM-15, Feb. 1967, pp. 52-60.

3. Tou, J. T.; and Heydorn, R. P.: Some Approaches to Optimum Feature Extraction. Proc. Second Sym. Computer and Information Sciences, ed. by J. T. Tou and R. Wilcox (Columbus, Ohio), Aug. 1966.

4. Chittineni, C. B.: On the Application of Divergence to Feature Selection in Pattern Recognition. IEEE Trans. Systems, Man, and Cybernetics, vol. SMC-2, no. 5, Nov. 1973, pp. 668-670.

5. Chittineni, C. B.: On Feature Extraction in Pattern Recognition. Internat. J. Information Sciences, vol. 6, 1973, pp. 191-200.

6. Chittineni, C. B.: On the Probability of Error and the Expected Bhattacharyya Distance in Multiclass Pattern Recognition. Proc. IEEE, vol. 60, no. 11, May 1972, pp. 1451-1452.

7. Bhattacharyya, A.: On a Measure of Divergence Between Two Statistical Populations Defined by Their Probability Distributions. Bull. Calcutta Math Soc., vol. 35, 1943, pp. 99-109.

8. Jeffreys, H.: An Invariant for the Prior Probability in Estimation Problems. Proc. Royal Stat. Soc. A, vol. 186, 1946, pp. 454-461.

9. Matusita, K.: On the Theory of Statistical Decision Functions. Ann. Instit. Stat. Math. (Tokyo), vol. 3, 1951, pp.17-35.

10. Matusita, K.: Classification Based on Distance in Multivariate Gaussian Case. Proc. 5th Berkeley Sym. Math. Stat. and Prob., vol. 1, 1967, pp. 299-304.

11. Kullback, S.; and Leibler, R. A.: On Information and Sufficiency. Ann. Math. Stat., vol. 22, 1951, pp.79-86.

12. Mahalanobis, P. C.: On the Generalized Distance in Statistics. Proc. India National Inst. of Science, vol. 2, 1936, pp. 49-55.

13. Fu, K. S.; and Min, P. J.: On Feature Selection in Multiclass Pattern Recognition. Tech. rpt. TR-EE68-17 (Purdue University), July 1968.

14. Beckenbach, E. F., ed: Applied Combinatorial Mathematics. John Wiley and Sons, Inc., New York, 1964.

15. Bodewig, E.: Matrix Calculus. Amsterdam: North Holland, 1959.