Distance-Based Approaches to Pattern Recognition via Embedding

Nicholas A. Nechval, *Member, IAENG,* Konstantin N. Nechval, Vadim Danovich, Gundars Berzins

***Abstract* — The most popular separation criterion of establishing rules for discrimination and recognition (classification) of patterns is the Fisher discriminant (separation) ratio. The approach proposed by Fisher assumes equality of population covariance matrices, but does not explicitly require multivariate normality. However, optimal classification performance of Fisher's discriminant function can only be expected when multivariate normality is present as well, since only good discrimination can ensure good allocation. In practice, we often are in need of analyzing input data samples, which are not adequate for Fisher's classification rule, such that the distributions of the groups are not multivariate normal or covariance matrices of those are different or there are strong multi-nonlinearities. In this paper, distance-based approaches for pattern classification (recognition) via embedding are proposed which allow one to classify, say, radar clutter into one of several major categories, including bird, weather, and target classes. These approaches do not require the arbitrary selection of priors as in the Bayesian classifier and represent the improved pattern recognition (classification) procedures that allows one to take into account the cases which are not adequate for Fisher's classification rule. Moreover, they allow one to classify sets of multivariate observations, where each of the sets contains more than one observation. For the cases, which are adequate for Fisher's classification rule, the proposed approaches give the results similar to that of Fisher's classification rule. For illustration, a numerical example is given.**

***Index Terms* — Pattern, embedding, classification, distance-based approaches**

I. Introduction

PATTERN recognition provides the solution to various problems from speech recognition, face recognition to classification of handwritten characters and medical diagnosis. The various application areas of pattern recognition are like bioinformatics, document classification, image analysis, data mining, industrial automation, biometric recognition, remote sensing, handwritten text analysis, medical diagnosis, speech recognition, statistics, mathematics, computer science, biology and many more. Similarity between all these applications is that for a solution-finding approach features have to be extracted and then analyzed for recognition and classification purpose. Three processes take place in pattern recognition task. First step is data acquisition. Data acquisition is the process of converting data from one form (speech, character, pictures etc.) into another form which should be acceptable to the computing device for further processing. Second step is data analysis. After data acquisition the task of analysis begins. During data analysis step the learning about the data takes place and information is collected about the different events and pattern classes available in the data. This information or knowledge about the data is used for further processing. Third step used for pattern recognition is classification. Its purpose is to decide the category of new data on the basis of knowledge received from data analysis process. There are many sub-problems in the design process. Many of these problems can indeed be solved. More complex learning, searching and optimization algorithms are developed with advances in computer technology. There remain many fascinating unsolved problems.

Pattern recognition aim is to classify data (patterns) based on either a priori knowledge or on statistical information extracted from the patterns. The patterns to be classified are usually groups of measurements or observations, defining points in an appropriate multidimensional space. Many pattern recognition methods can be decomposed into two stages: discrimination followed by classification. In some cases, the decomposition is explicit while in others it is a matter of interpretation. Discrimination and classification represent multivariate techniques concerned with separating distinct sets of objects (or observations) and allocating new objects (observations) to previously defined groups. There exist situations in which one may interested in (1) discrimination: separating, say, two classes of objects or (2) classification: assigning a new object to one of two classes (or both).

The most popular separation criterion of establishing rules for discrimination and classification of patterns is the Fisher discriminant (separation) ratio. Fisher's idea was to transform the ($p \geq 2$) multivariate observations $\mathbf{y}$ to univariate observations z such that the z's derived from populations π_1 and π_2 were separated as much as possible. Fisher suggested taking linear combinations of $\mathbf{y}$ to create z's because they are simple enough functions of the $\mathbf{y}$ to be handled easily. Fisher's approach does not assume that the populations are normal. It does, however, implicitly assume that the population covariance matrices are equal, because a pooled estimate of the common covariance matrix is used.

Manuscript received March 23, 2014; revised April 10, 2014. This research was supported in part by Grant No. 09.1544 from the Latvian Council of Science and the National Institute of Mathematics and Informatics of Latvia.

Nicholas. A. Nechval is with the Statistics Department, EVF Research Institute, University of Latvia, Riga LV-1050, Latvia (e-mail: nechval@junik.lv).

Konstantin N. Nechval is with the Applied Mathematics Department, Transport and Telecommunication Instutute, Riga LV-1019, Latvia (e-mail: konstan@tsi.lv).

Vadim Danovich is with the Cybernetics Department, University of Latvia, Riga LV-1050, Latvia (e-mail: vadims@lu.lv).

Gundars Berzins is with the Cybernetics Department, University of Latvia, Riga LV-1050, Latvia (e-mail: gundars.berzins@lu.lv).

ISBN: 978-988-19253-5-0
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Fisher's linear discriminant analysis has been successfully used as dimensionality reduction technique to many classification problems, such as face recognition and multimedia information retrieval. The Fisher discriminant criterion is the benchmark for the linear discrimination in multidimensional space [1]. The criterion purpose of the Fisher linear discriminant for pattern analysis is to find an optimal discriminant direction based on the Fisher criterion so that the projected set of training samples on it has the maximal ratio of between-class distance to within-class distance [2]. Sammon extended the Fisher linear discriminant method to the optimal discriminant plane in 1970 [3]. Then Foley and Sammon [4] further extended this in 1975 and proposed the optimal set of discriminant vectors by which the well-known Foley-Sammon Transform (FST) can be constituted. Their important result has attracted many researchers' attention in the field of pattern recognition and has been used in many pattern classification applications.

II. Pattern Classification Problem

The classification problem consists in the following. There are m classes (populations), the elements (objects) of which are characterized by p measurements (features). Next, suppose that we are investigating a certain object on the basis of the corresponding p measurements. We postulate that this object can be regarded as a "random drawing" from one of the m populations but we do not know from which one. We suppose that m samples are available, each sample being drawn from a different class (population). The elements of these samples are realizations of p-dimensional random variables. After a sample of p-dimensional vectors of observations on the object is drawn from a class known a priori to be one of the above set of m classes, the problem is to infer from which class the sample has been drawn. The decision rule should be in the form of associating the sample of observations on the object with one of the m samples and declaring that the object has come from the same class as the sample with which it is associated.

Classification is often referred to simply as discriminant analysis. In engineering and computer science, classification is usually called pattern recognition. Some writers use the term classification analysis to describe cluster analysis, in which the observations are clustered according to variable values rather than into predefined classes.

In classification, a sampling unit (subject or object) whose class membership is unknown is assigned to a class on the basis of the vector of p measured values, $\mathbf{y}$, associated with the unit. To classify the unit, we must have available a previously obtained sample of observation vectors from each class. Then one approach is to compare $\mathbf{y}$ with the mean vectors $\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2, ..., \bar{\mathbf{y}}_k$ of the k samples and assign the unit to the class whose $\bar{\mathbf{y}}_i$ is closest to $\mathbf{y}$.

III. Fisher's Approach to Pattern Classification into Two Classes

When there are two populations (classes), we can use a classification procedure due to Fisher [1]. The principal assumption for Fisher's procedure is that the two populations have the same covariance matrix ($\Sigma_1 = \Sigma_2$). Normality is not required. We obtain a sample from each of the two populations and compute $\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2$, and $\mathbf{S}_{12}$. A simple procedure for classification into one of the two classes denoted by C_1 and C_2 can be based on the discriminant function,

$$z = \mathbf{w}'\mathbf{y} = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\mathbf{S}_{12}^{-1}\mathbf{y}, \tag{1}$$

where $\mathbf{y}$ is the vector of measurements on a new sampling unit that we wish to classify into one of the two classes (populations), $\mathbf{w}$ is a direction which is determined from maximization of the ratio of between-class to within-class variances proposed by Fisher,

$$J_F = \frac{[\mathbf{w}'(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)]^2}{\mathbf{w}'\mathbf{S}_{12}\mathbf{w}}, \tag{2}$$

$\mathbf{S}_{12}$ is the pooled within-class covariance matrix, in its bias-corrected form given by

$$\mathbf{S}_{12} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}, \tag{3}$$

$\mathbf{S}_1$ and $\mathbf{S}_2$ are the unbiased estimates of the covariance matrices of classes C_1 and C_2, respectively, and there are n_i observations in class C_i (n_1+n_2=n). The solution for $\mathbf{w}$ that maximizes J_F can be obtained by differentiating J_F with respect to $\mathbf{w}$ and equating to zero. This yields

$$\frac{2\mathbf{w}'(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)}{\mathbf{w}'\mathbf{S}_{12}\mathbf{w}}\left[(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2) - \left(\frac{\mathbf{w}'(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)}{\mathbf{w}'\mathbf{S}_{12}\mathbf{w}}\right)\mathbf{S}_{12}\mathbf{w}\right] = 0. \tag{4}$$

Since we are interested in the direction of $\mathbf{w}$ (and noting that $\mathbf{w}'(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2) / \mathbf{w}'\mathbf{S}_{12}\mathbf{w}$ is a scalar), we must have

$$\mathbf{w} \propto \mathbf{S}_{12}^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2). \tag{5}$$

We may take equality without loss of generality. For convenience we speak of classifying $\mathbf{y}$ rather than classifying the subject or object associated with $\mathbf{y}$.

To determine whether $\mathbf{y}$ is closer to $\bar{\mathbf{y}}_1$ or $\bar{\mathbf{y}}_2$, we check to see if z in (1) is closer to the transformed mean $\bar{z}_1$ or to $\bar{z}_2$, where

$$\bar{z}_1 = \mathbf{w}'\bar{\mathbf{y}}_1 = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\mathbf{S}_{12}^{-1}\bar{\mathbf{y}}_1, \tag{6}$$

$$\bar{z}_2 = \mathbf{w}'\bar{\mathbf{y}}_2 = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\mathbf{S}_{12}^{-1}\bar{\mathbf{y}}_2. \tag{7}$$

Fisher's linear classification procedure [1] assigns $\mathbf{y}$ to C_1 if $z = \mathbf{w}'\mathbf{y}$ is closer to $\bar{z}_1$ than to $\bar{z}_2$ and assigns $\mathbf{y}$ to C_2 if z is closer to $\bar{z}_2$. It will be noted that z is closer to $\bar{z}_1$ if

$$z > \frac{\bar{z}_1 + \bar{z}_2}{2}. \tag{8}$$

This is true in general because $\bar{z}_1$ is always greater than $\bar{z}_2$, which can easily be shown as follows:

$$\bar{z}_1 - \bar{z}_2 = \mathbf{w}'(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2) = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\mathbf{S}_{12}^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2) > 0, \tag{9}$$

because $\mathbf{S}_{12}^{-1}$ is positive definite. Thus $\bar{z}_1 > \bar{z}_2$. [If $\mathbf{w}$ were of

ISBN: 978-988-19253-5-0
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

the form $\mathbf{w}' = (\bar{\mathbf{y}}_2 - \bar{\mathbf{y}}_1)'\mathbf{S}_{12}^{-1}$, then $\bar{z}_2 - \bar{z}_1$ would be positive.] Since $(\bar{z}_1 + \bar{z}_2)/2$ is the midpoint, $z > (\bar{z}_1 + \bar{z}_2)/2$ implies that z is closer to $\bar{z}_1$. By (9) the distance from $\bar{z}_1$ to $\bar{z}_2$ is the same as that from $\bar{\mathbf{y}}_1$ to $\bar{\mathbf{y}}_2$.

To express the classification rule in terms of **y**, we first write $(\bar{z}_1 + \bar{z}_2)/2$ in the form

$$\frac{\bar{z}_1 + \bar{z}_2}{2} = \frac{\mathbf{w}'(\bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_2)}{2} = \frac{(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\mathbf{S}_{12}^{-1}(\bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_2)}{2}. \quad (10)$$

Then the classification rule becomes: Assign **y** to C_1 if

$$\mathbf{w}'\mathbf{y} = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)\mathbf{S}_{12}^{-1}\mathbf{y} > \frac{(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\mathbf{S}_{12}^{-1}(\bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_2)}{2}, \quad (11)$$

and assign **y** to C_2 if

$$\mathbf{w}'\mathbf{y} = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)\mathbf{S}_{12}^{-1}\mathbf{y} < \frac{(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\mathbf{S}_{12}^{-1}(\bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_2)}{2}. \quad (12)$$

Fisher's approach [1] using (11) and (12) is essentially nonparametric because no distributional assumptions were made. However, if the two populations are normal with equal covariance matrices, then this method is (asymptotically) optimal; that is, the probability of misclassification is minimized.

IV. Approaches to Pattern Classification into Two Classes via Embedding

A. Classification Based on Mahalanobis Distance

Let us assume that the two populations have the same covariance matrix ($\Sigma_1 = \Sigma_2$).

If **y** is embedded in the sample from C_1, the Mahalanobis distance between two mean vectors $\bar{\mathbf{y}}_{\bullet 1}$ and $\bar{\mathbf{y}}_2$ is given by

$$d_{\bullet 12} = (\bar{\mathbf{y}}_{\bullet 1} - \bar{\mathbf{y}}_2)'\mathbf{S}_{\bullet 12}^{-1}(\bar{\mathbf{y}}_{\bullet 1} - \bar{\mathbf{y}}_2). \quad (13)$$

If **y** is embedded in the sample from C_2, the Mahalanobis distance between two mean vectors $\bar{\mathbf{y}}_1$ and $\bar{\mathbf{y}}_{2\bullet}$ is given by

$$d_{12\bullet} = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_{2\bullet})'\mathbf{S}_{12\bullet}^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_{2\bullet}). \quad (14)$$

Then the classification rule becomes: Assign **y** to C_1 if

$$d_{\bullet 12} > d_{12\bullet}, \quad (15)$$

and assign **y** to C_2 if

$$d_{12\bullet} > d_{\bullet 12}. \quad (16)$$

If $(\Sigma_1 = \Sigma_2)$ does not hold, then instead of $\mathbf{S}_{12}$ we use

$$\mathbf{S}_{12}^{\circ} = \frac{\mathbf{S}_1}{n_1} + \frac{\mathbf{S}_2}{n_2}. \quad (17)$$

Remark. If $n_1 = n_2 = n$, then

$$\frac{n-1}{n+n-2} = \frac{1}{2}, \quad (18)$$

so

$$\mathbf{S}_{12}^{\circ} = \frac{\mathbf{S}_1}{n_1} + \frac{\mathbf{S}_2}{n_2} = \frac{(n_1-1)\mathbf{S}_1 + (n_2-1)\mathbf{S}_2}{n_1+n_2-2}\left(\frac{1}{n}+\frac{1}{n}\right) = \mathbf{S}_{12}\left(\frac{1}{n}+\frac{1}{n}\right). \quad (19)$$

With equal sample sizes, the large sample procedure is essentially the same as the procedure based on the pooled covariance matrix.

B. Classification Based on Generalized Euclidean Distance

Let us assume that the two populations have the same covariance matrix ($\Sigma_1 = \Sigma_2$).

If **y** is embedded in the sample from C_1, the generalized Euclidean distance (squared) between two mean vectors $\bar{\mathbf{y}}_{\bullet 1}$ and $\bar{\mathbf{y}}_2$ is given by

$$\tilde{d}_{\bullet 12} = \frac{(\bar{\mathbf{y}}_{\bullet 1} - \bar{\mathbf{y}}_2)'(\bar{\mathbf{y}}_{\bullet 1} - \bar{\mathbf{y}}_2)}{|\mathbf{S}_{\bullet 12}|}. \quad (20)$$

If **y** is embedded in the sample from C_2, the generalized Euclidean distance between two mean vectors $\bar{\mathbf{y}}_1$ and $\bar{\mathbf{y}}_{2\bullet}$ is given by

$$\tilde{d}_{12\bullet} = \frac{(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_{2\bullet})'(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_{2\bullet})}{|\mathbf{S}_{12\bullet}|}. \quad (21)$$

Then the classification rule becomes: Assign **y** to C_1 if

$$\tilde{d}_{\bullet 12} > \tilde{d}_{12\bullet}, \quad (22)$$

and assign **y** to C_2 if

$$\tilde{d}_{12\bullet} > \tilde{d}_{\bullet 12}. \quad (23)$$

If $(\Sigma_1 = \Sigma_2)$ does not hold, then instead of $\mathbf{S}_{12}$ we use (17).

C. Classification Based on Modified Euclidean Distance

Let us assume that the two populations have the same covariance matrix ($\Sigma_1 = \Sigma_2$).

If **y** is embedded in the sample from C_1, the modified Euclidean distance between two mean vectors $\bar{\mathbf{y}}_{\bullet 1}$ and $\bar{\mathbf{y}}_{12}$ is given by

$$\breve{d}_{\bullet 1} = \frac{(\bar{\mathbf{y}}_{\bullet 1} - \bar{\mathbf{y}}_{12})'(\bar{\mathbf{y}}_{\bullet 1} - \bar{\mathbf{y}}_{12})}{|\mathbf{S}_{\bullet 12}|}, \quad (24)$$

where

$$\bar{\mathbf{y}}_{12} = \sum_{i=1}^{2} n_i \bar{\mathbf{y}}_i \Big/ \sum_{i=1}^{2} n_i. \quad (25)$$

If **y** is embedded in the sample from C_2, the generalized Euclidean distance between two mean vectors $\bar{\mathbf{y}}_{2\bullet}$ and $\bar{\mathbf{y}}_{12}$ is given by

$$\breve{d}_{2\bullet} = \frac{(\bar{\mathbf{y}}_{2\bullet} - \bar{\mathbf{y}}_{12})'(\bar{\mathbf{y}}_{2\bullet} - \bar{\mathbf{y}}_{12})}{|\mathbf{S}_{12\bullet}|}. \quad (26)$$

Then the classification rule becomes: Assign **y** to C_1 if

$$\breve{d}_{\bullet 1} + \breve{d}_2 > \breve{d}_1 + \breve{d}_{2\bullet}, \quad (27)$$

and assign **y** to C_2 if

$$\breve{d}_1 + \breve{d}_{2\bullet} > \breve{d}_{\bullet 1} + \breve{d}_2. \quad (28)$$

ISBN: 978-988-19253-5-0
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

If ($\Sigma_1 = \Sigma_2$) does not hold, then instead of $\mathbf{S}_{12}$ we use (17).

V. Known Approaches to Pattern Classification into Several Classes

A. Classification Based on Mahalanobis Distance

Equal Population Covariance Matrices. In this section we discuss classification rules for several classes. As in the two-class case, we use a sample from each of the k classes to find the sample mean vectors $\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2, ..., \bar{\mathbf{y}}_k$. For a vector $\mathbf{y}$ whose class membership is unknown, one approach is to use a distance function to find the mean vector that $\mathbf{y}$ is closest to and assign $\mathbf{y}$ to the corresponding class.

We assume $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k$ and can estimate the common population covariance matrix by a pooled sample covariance matrix

$$\mathbf{S}_{\text{pl}} = \sum_{i=1}^{k}(n_i - 1)\mathbf{S}_i \left[\sum_{i=1}^{k} n_i - k\right]^{-1}, \tag{29}$$

where n_i and $\mathbf{S}_i$ are the sample size and covariance matrix of the ith class. We compare $\mathbf{y}$ to each $\bar{\mathbf{y}}_i$, i=1, 2, ..., k, by the distance function

$$D_i^2(\mathbf{y}) = (\mathbf{y} - \bar{\mathbf{y}}_i)'\mathbf{S}_{\text{pl}}^{-1}(\mathbf{y} - \bar{\mathbf{y}}_i) \tag{30}$$

and assign $\mathbf{y}$ to the class for which D_i^2 is smallest. This classification rule is based on the assumption $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k$. The resulting classification rules are sensitive to heterogeneity of covariance matrices. Observations tend to be classified too frequently into classes whose covariance matrices have larger variances on the diagonal. Thus, the population covariance matrices should not be assumed to be equal if there is reason to suspect otherwise.

Unequal Population Covariance Matrices. If $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k$ does not hold, the classification rules can easily be altered to preserve optimality of classification rates. In place of (30), we can use

$$D_i^2(\mathbf{y}) = (\mathbf{y} - \bar{\mathbf{y}}_i)'\mathbf{S}_i^{-1}(\mathbf{y} - \bar{\mathbf{y}}_i), \quad i=1, 2, ..., k, \tag{31}$$

where $\mathbf{S}_i$ is the sample covariance matrix for the ith class. As before, we would assign $\mathbf{y}$ to the class for which $D_i^2(\mathbf{y})$ is smallest (with $\mathbf{S}_i$ in place of $\mathbf{S}_{\text{pl}}$).

VI. Approaches to Pattern Classification into Several Classes via Embedding

A. Classification Based on Total Mahalanobis Distance

Equal Population Covariance Matrices. Let us assume that each of the k populations has the same covariance matrix ($\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k$). The Mahalanobis distance between two mean vectors $\bar{\mathbf{y}}_i$ and $\bar{\mathbf{y}}_j$, where $i, j \in \{1, 2, \ldots, k\}$, $i \neq j$, is given by

$$d_{ij} = (\bar{\mathbf{y}}_i - \bar{\mathbf{y}}_j)'\mathbf{S}_{ij}^{-1}(\bar{\mathbf{y}}_i - \bar{\mathbf{y}}_j). \tag{32}$$

If $\mathbf{y}$ has been embedded in the sample from C_i, then the Mahalanobis distance between two vectors $\bar{\mathbf{y}}_{\bullet i}$ and $\bar{\mathbf{y}}_j$ is given by

$$d_{\bullet ij} = (\bar{\mathbf{y}}_{\bullet i} - \bar{\mathbf{y}}_j)'\mathbf{S}_{\bullet ij}^{-1}(\bar{\mathbf{y}}_{\bullet i} - \bar{\mathbf{y}}_j). \tag{33}$$

If $\mathbf{y}$ has been embedded in the sample from C_j, then the Mahalanobis distance between two vectors $\bar{\mathbf{y}}_i$ and $\bar{\mathbf{y}}_{j\bullet}$ is given by

$$d_{ij\bullet} = (\bar{\mathbf{y}}_i - \bar{\mathbf{y}}_{j\bullet})'\mathbf{S}_{ij\bullet}^{-1}(\bar{\mathbf{y}}_i - \bar{\mathbf{y}}_{j\bullet}). \tag{34}$$

Let

$$d_r(\mathbf{y}) = \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} d_{ij}, \ r \in \{1, 2, \ldots, k\} \tag{35}$$

be the total Mahalanobis distance in the case of pattern classification into k classes, where

$$d_{ij} = d_{\bullet ij}, \text{ if } i = r, \tag{36}$$

and

$$d_{ij} = d_{ij\bullet}, \text{ if } j = r. \tag{37}$$

Then the classification rule becomes: Assign $\mathbf{y}$ to the class C_r, $r \in \{1, 2, \ldots, k\}$, for which $d_r(\mathbf{y})$ is largest.

Unequal Population Covariance Matrices. If ($\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k$) does not hold, then instead of $\mathbf{S}_{ij}$ we use

$$\mathbf{S}_{ij}^{\circ} = \frac{\mathbf{S}_i}{n_i} + \frac{\mathbf{S}_j}{n_j}. \tag{38}$$

B. Classification Based on Total Generalized Euclidean Distance

Equal Population Covariance Matrices. Let us assume that each of the k populations has the same covariance matrix ($\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k$). The generalized Euclidean distance between two mean vectors $\bar{\mathbf{y}}_i$ and $\bar{\mathbf{y}}_j$, where $i, j \in \{1, 2, \ldots, k\}$, $i \neq j$, is given by

$$\tilde{d}_{ij} = \frac{(\bar{\mathbf{y}}_i - \bar{\mathbf{y}}_j)'(\bar{\mathbf{y}}_i - \bar{\mathbf{y}}_j)}{|\mathbf{S}_{\text{pl}}|}. \tag{39}$$

If $\mathbf{y}$ is embedded in the sample from C_i, the generalized Euclidean distance between two vectors $\bar{\mathbf{y}}_{\bullet i}$ and $\bar{\mathbf{y}}_j$ is given by

$$\tilde{d}_{\bullet ij} = \frac{(\bar{\mathbf{y}}_{\bullet i} - \bar{\mathbf{y}}_j)'(\bar{\mathbf{y}}_{\bullet i} - \bar{\mathbf{y}}_j)}{|\mathbf{S}_{\text{pl}(\bullet i)}|}. \tag{40}$$

If $\mathbf{y}$ is embedded in the sample from C_j, then the generalized Euclidean distance between two vectors $\bar{\mathbf{y}}_i$ and $\bar{\mathbf{y}}_{j\bullet}$ is given by

$$\tilde{d}_{ij\bullet} = \frac{(\bar{\mathbf{y}}_i - \bar{\mathbf{y}}_{j\bullet})'(\bar{\mathbf{y}}_i - \bar{\mathbf{y}}_{j\bullet})}{|\mathbf{S}_{\text{pl}(j\bullet)}|}. \tag{41}$$

Let

$$\tilde{d}_r(\mathbf{y}) = \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} \tilde{d}_{ij}, \ r \in \{1, 2, \ldots, k\}, \tag{42}$$

be the total generalized Euclidean distance in the case of pattern classification into k classes, where

ISBN: 978-988-19253-5-0
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

$$\tilde{d}_{ij} = \tilde{d}_{\bullet ij}, \text{ if } i = r, \tag{43}$$

and

$$\tilde{d}_{ij} = \tilde{d}_{ij\bullet}, \text{ if } j = r. \tag{44}$$

Then the classification rule becomes: Assign **y** to the class C_r, $r \in \{1, 2, \ldots, k\}$, for which $\tilde{d}_r(\mathbf{y})$ is largest.

Unequal Population Covariance Matrices. If $(\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k)$ does not hold, then instead of $\mathbf{S}_{\text{pl}}$ we use

$$\mathbf{S}^{\circ} = \sum_{i=1}^{k} \frac{\mathbf{S}_i}{n_i}. \tag{45}$$

C. Classification Based on Total Modified Euclidean Distance

Equal Population Covariance Matrices. Let us assume that each of the k populations has the same covariance matrix $(\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k)$. The modified Euclidean distance between two mean vectors $\bar{\mathbf{y}}_i$ and $\bar{\mathbf{y}}$, $i \in \{1, 2, \ldots, k\}$, is given by

$$\breve{d}_i = \frac{(\bar{\mathbf{y}}_i - \bar{\mathbf{y}})'(\bar{\mathbf{y}}_i - \bar{\mathbf{y}})}{|\mathbf{S}_{\text{pl}}|}, \tag{46}$$

where

$$\bar{\mathbf{y}} = \sum_{i=1}^{k} n_i \bar{\mathbf{y}}_i \Big/ \sum_{i=1}^{k} n_i \tag{47}$$

represents the 'overall average'.

If **y** has been embedded in the sample from C_i, then the modified Euclidean distance between two vectors $\bar{\mathbf{y}}_{\bullet i}$ and $\bar{\mathbf{y}}$ is given by

$$\breve{d}_{\bullet i} = \frac{(\bar{\mathbf{y}}_{\bullet i} - \bar{\mathbf{y}})'(\bar{\mathbf{y}}_{\bullet i} - \bar{\mathbf{y}})}{|\mathbf{S}_{\text{pl}(\bullet i)}|}, \; i \in \{1, 2, \ldots, k\}. \tag{48}$$

Let

$$\breve{d}_r(\mathbf{y}) = \sum_{i=1}^{k} \breve{d}_i, \; r \in \{1, 2, \ldots, k\}, \tag{49}$$

be the total modified Euclidean distance in the case of pattern classification into k classes, where

$$\breve{d}_i = \breve{d}_{\bullet i}, \text{ if } i = r. \tag{50}$$

Then the classification rule becomes: Assign **y** to to the class C_r, $r \in \{1, 2, \ldots, k\}$, for which $\breve{d}_r(\mathbf{y})$ is largest.

Unequal Population Covariance Matrices. If $(\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k)$ does not hold, then instead of $\mathbf{S}_{\text{pl}}$ we use (45).

VII. Illustrative Example of Pattern Classification

Consider the observations on p=2 variables from k=3 populations (classes) [5]. The input data samples are given below.

$$\begin{matrix} (n_1{=}3) & & (n_2{=}3) & & (n_3{=}3) \end{matrix}$$

$$C_1 = \begin{bmatrix} -2 & 5 \\ 0 & 3 \\ -1 & 1 \end{bmatrix}; \quad C_2 = \begin{bmatrix} 0 & 6 \\ 2 & 4 \\ 1 & 2 \end{bmatrix}; \quad C_3 = \begin{bmatrix} 1 & -2 \\ 0 & 0 \\ -1 & -4 \end{bmatrix}. \tag{51}$$

We found that

$$\bar{\mathbf{y}}_1 = \begin{bmatrix} -1 \\ 3 \end{bmatrix}, \quad \bar{\mathbf{y}}_2 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}, \quad \bar{\mathbf{y}}_3 = \begin{bmatrix} 0 \\ -2 \end{bmatrix}, \quad \bar{\mathbf{y}} = \begin{bmatrix} 0 \\ 5/3 \end{bmatrix}. \tag{52}$$

$$\mathbf{S}_{\text{pl}} = \begin{bmatrix} 1 & -0.33333 \\ -0.33333 & 4 \end{bmatrix}. \tag{53}$$

Suppose that we have to classify the new observation $\mathbf{y}'$=[1, 3] into the above classes. Let us assume that each of the k=3 populations has the same covariance matrix $(\Sigma_1 = \Sigma_2 = \Sigma_3)$.

Classification Based on Total Mahalanobis Distance. It follows from (35) that

$$d_1(\mathbf{y}) = 20.86, \quad d_2(\mathbf{y}) = 26.33, \quad d_3(\mathbf{y}) = 15.49. \tag{54}$$

Thus, since

$$d_2(\mathbf{y}) = \max_{r \in \{1, 2, 3\}} d_r(\mathbf{y}), \tag{55}$$

we assign **y** to class C_2.

Classification Based on Total Generalized Euclidean Distance. It follows from (42) that

$$\tilde{d}_1(\mathbf{y}) = 15.14, \quad \tilde{d}_2(\mathbf{y}) = 21.91, \quad \tilde{d}_3(\mathbf{y}) = 7.51. \tag{56}$$

Thus, since

$$\tilde{d}_2(\mathbf{y}) = \max_{r \in \{1, 2, 3\}} \tilde{d}_r(\mathbf{y}), \tag{57}$$

we assign **y** to class C_2.

Classification Based on Total Modified Euclidean Distance. It follows from (49) that

$$\breve{d}_1(\mathbf{y}) = 5.066, \quad \breve{d}_2(\mathbf{y}) = 7.312, \quad \breve{d}_3(\mathbf{y}) = 2.596. \tag{58}$$

Thus, since

$$\breve{d}_2(\mathbf{y}) = \max_{r \in \{1, 2, 3\}} \breve{d}_r(\mathbf{y}), \tag{59}$$

we assign **y** to class C_2.

It will be noted that the procedures proposed in this paper give the same result in the above case that of Fisher's procedure which was used in [5].

VIII. Conclusion and Future Work

Linear discriminants may be used to discriminate any number of classes of patterns, but are perhaps most commonly used when there are only two classes. An example of such a problem is in detection, where it is required that a target pattern, such as a vehicle in a radar image, is detected from among the uninteresting background patterns. Many detection problems are specified so that the classifier must produce either a particular detection rate or an upper bound for the rate at which false detections are produced. Each of these specifications will be referred to as an "operating point" for the classifier. The Fisher discriminant [1] is the benchmark for the linear discrimination between two classes in multidimensional space. It is extremely quick to calculate since it is based only on the first and second moments of each distribution. Also, it may be shown to maximize a measure of the separation which is not specific to a particular distribution type. This

ISBN: 978-988-19253-5-0
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

makes the Fisher discriminant extremely robust.

Since many fault diagnosis problems can be considered as a multi-class classification problems, pattern recognition methods with good generalization and accurate performances have been proposed in recent years. Choi et al. [6] proposed a fault detection and isolation methodology based on principal component analysis–Gaussian mixture model and discriminant analysis–Gaussian mixture model. Fisher's linear discriminant analysis (FLDA) has been proved to outperform the principal component analysis (PCA) in discriminating different classes, in the aspect that PCA aims at reconstruction instead of classification, while FLDA seeks directions that are optimal for discrimination [7]. Fisher's linear discriminant analysis is a widely used multivariate statistical technique with two closely related goals: discrimination and classification. The technique is very popular among users of discriminant analysis. Some of the reasons for this are its simplicity and unnecessity of strict assumptions. In its original form, proposed by Fisher, the method assumes equality of population covariance matrices, but does not explicitly require multivariate normality. However, optimal classification performance of Fisher's discriminant function can only be expected when multivariate normality is present as well, since only good discrimination can ensure good allocation.

In practice, we often are in need of analyzing input data samples, which are not adequate for Fisher's classification rule, such that the distributions of the groups are not multivariate normal or covariance matrices of those are different or there are strong multi-nonlinearities. In particular, the situation of pattern classification, which sometimes produces non-linear separation of classes, is not adequate for Fisher's classification rule. One solution to address this problem would be the application of kernels to input data, which would essentially transform the input to a higher dimensional space, wherein the probability of linearly separating the classes is higher. In kernel methods, there is a strong possibility that the higher dimensional space may be non-linear in nature as opposed to the linear input space; and the separation of classes would be linear in feature space and non-linear in input space. Any kernel-based method employed for classification, includes two major steps [8, 9]: (i) mapping the non-linearly separable input data from its current lower dimensional space (*input space*) to a linearly separable data in the (*feature space*), and (ii) classifying patterns in the feature space. The use of kernels would become clear from the illustration shown in Fig. 1.

Assume that the input data is of the form,

$$\mathbf{y} = (y_1, y_2, y_3, ..., y_m)'. \tag{60}$$

Then, using a function F, we can map the input data set $\mathbf{y}$ into a higher dimensional space as,

$$\mathbf{y} = (y_1, y_2, y_3, ..., y_m)' \rightarrow \mathbf{F}(\mathbf{y}) = (F(y_1), F(y_2), F(y_3), ..., F(y_m)'. \tag{61}$$

This is essentially same as mapping the input space $\mathscr{Y}$ into a new space $\mathscr{F}$. The space $\mathscr{F}$ is known as the Feature Space. It also takes into consideration the input features introduced in order to classify the data to the correct category label.

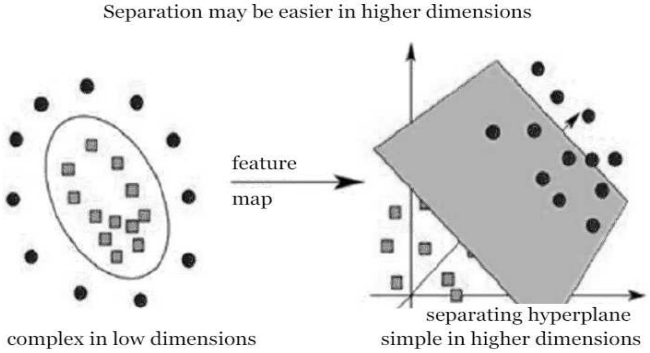

Fig. 1. Dimension transformation.

This paper proposes the improved approaches to pattern classification which represent the new distance-based embedding procedures that allow one to take into account the cases which are not adequate for Fisher's classification rule. Moreover, these approaches allow one to classify sets of multivariate observations, where each of the sets contains more than one observation. For the cases, which are adequate for Fisher's classification rule, the proposed approaches give the results similar to that of FLDA.

The methodology described here can be extended in several different directions to handle various problems of pattern classification (recognition) that arise in practice (in particular, the problem of changepoint detection in a sequence of multivariate observations).

ACKNOWLEDGMENT

This research was supported in part by Grant No. 06.1936, Grant No. 07.2036, Grant No. 09.1014, and Grant No. 09.1544 from the Latvian Council of Science and the National Institute of Mathematics and Informatics of Latvia.

REFERENCES

[1] R. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, pp.178–188, 1936.

[2] Q. Tian, Y. Fainman, Z. H. Gu, and S. Lee, "Comparison of statistical pattern-recognition algorithms for hybrid processing, I. linear-mapping algorithm," *J. Opt. Soc. Am. A.*, vol. 5, pp. 1655–1669, 1988.

[3] J. W. Sammon, "An optimal discriminant plane," *IEEE Trans. Computer*, vol. C-19, pp. 826–829, 1970.

[4] D. Foley, J. Sammon, "An optimal set of discriminant vectors," *IEEE Trans. Computer*, vol. C-24, pp. 281–289, 1975.

[5] N. A. Nechval, M. Purgailis, D. Skiltere, and K. N. Nechval, "Pattern recognition based on comparison of Fisher's maximum separations," *Proceedings of the 7th International Conference on Neural Networks and Artificial Intelligence (ICNNAI'2012)*. Minsk, Belarus, pp. 65–69, 2012.

[6] S. W. Choi, J. H. Park,. and I. B. Lee, "Process monitoring using a Gaussian mixture model via principal component analysis and discriminant analysis," *Computers and Chemical Engineering*, vol. 28, pp. 1377–1387, 2004.

[7] L. Chiang, E. L. Russell, and R. D. Braatz, "Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares, and principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 50, pp. 243–252, 2000.

[8] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, 2000.

[9] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.