Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 41 (2008) 3138–3152
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

Linear dimensionality reduction by maximizing the Chernoff distance in the transformed space

Luis Rueda[a,*,1], Myriam Herrera[b]

[a]*Department of Computer Science, University of Concepción, Edmundo Larenas 215, Concepción 4030000, Chile*
[b]*Institute of Informatics, National University of San Juan, Cereceto y Meglioli, San Juan 5400, Argentina*

Received 11 August 2006; received in revised form 17 September 2007; accepted 12 January 2008

Abstract

Linear dimensionality reduction (LDR) techniques are quite important in pattern recognition due to their linear time complexity and simplicity. In this paper, we present a novel LDR technique which, though linear, aims to maximize the Chernoff distance in the transformed space; thus, augmenting the class separability in such a space. We present the corresponding criterion, which is maximized via a gradient-based algorithm, and provide convergence and initialization proofs. We have performed a comprehensive performance analysis of our method combined with two well-known classifiers, linear and quadratic, on synthetic and real-life data, and compared it with other LDR techniques. The results on synthetic and standard real-life data sets show that the proposed criterion outperforms the latter when combined with both linear and quadratic classifiers.

Keywords: Linear dimensionality reduction; Pattern classification; Discriminant analysis

1. Introduction

Linear dimensionality reduction (LDR) techniques have been studied for a long time in the field of pattern recognition. They are typically the preferred ones due to their efficiency, and because they are simpler to implement and understand. We assume that we are dealing with two classes, ω_1 and ω_2, which are represented by two normally distributed n-dimensional random vectors, $\mathbf{x}_1 \sim N(\mathbf{m}_1, \mathbf{S}_1)$ and $\mathbf{x}_2 \sim N(\mathbf{m}_2, \mathbf{S}_2)$, and whose *a priori* probabilities are p_1 and p_2, respectively. The aim is to linearly transform $\mathbf{x}_1$ and $\mathbf{x}_2$ into new normally distributed random vectors $\mathbf{y}_1 = \mathbf{A}\mathbf{x}_1$ and $\mathbf{y}_2 = \mathbf{A}\mathbf{x}_2$ of dimension d, $d < n$, using a matrix $\mathbf{A}$ of order $d \times n$, in such a way that the classification error in the transformed space is as small as possible.

1.1. Related work

Various schemes that yield LDR have been reported in the literature, including the well known *Fisher's discriminant* (FD) approach [1], and its extensions: the *direct* FD analysis [2], its kernelized version for face recognition [3], the combined principal component analysis (PCA) and linear discriminant analysis (LDA) [4], the kernelized PCA and LDA [5], and a two-dimensional FD-based approach for face recognition [6]. An improvement to the FD approach that decomposes classes into subclasses has been proposed in Ref. [7]. Rueda et al. [8] showed that the optimal classifier between two normally distributed classes can be linear even when the covariance matrices *are not equal*. In Ref. [9], a new approach to selecting the *best hyperplane classifier* (BHC), which is obtained from the optimal pairwise linear classifier, has been introduced. A computationally intensive method for LDR was proposed in Ref. [10], which aims to minimize the classification error in the transformed space and operates by computing (or approximating) the *exact* values for the integrals. This approach, though extremely time consuming, does not guarantees an optimal LDR. Another criterion used for dimensionality reduction is the subclass discriminant analysis [11], which aims to optimally divide the classes into subclasses, and then performs the reduction followed by classification.

We now focus on two LDR approaches which are closely related to our proposed method. Let $\mathbf{S}_W = p_1\mathbf{S}_1 + p_2\mathbf{S}_2$ and

* Corresponding author. Tel.: +56 41 220 4305; fax: +56 41 222 1770.
E-mail addresses: lrueda@inf.udec.cl (L. Rueda), mherrera@iinfo.unsj.edu.ar (M. Herrera).
[1] Partially supported by the Chilean National Fund for Scientific and Technological Development, FONDECYT Grant no. 1060904.

0031-3203/$30.00
doi:10.1016/j.patcog.2008.01.016

$\mathbf{S}_E = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$ be the within-class and between-class scatter matrices, respectively. The well-known FD criterion consists of maximizing the Mahalanobis distance between the transformed distributions by finding $\mathbf{A}$ that maximizes the following function [1]:

$$J_{FD}(\mathbf{A}) = \mathrm{tr}\{(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1}(\mathbf{A}\mathbf{S}_E\mathbf{A}^t)\}. \quad (1)$$

The matrix $\mathbf{A}$ that maximizes (1) is obtained by finding the eigenvalue decomposition of the matrix:

$$\mathbf{S}_{FD} = \mathbf{S}_W^{-1}\mathbf{S}_E, \quad (2)$$

and taking the d eigenvectors whose eigenvalues are the largest ones. Since $\mathbf{S}_E$ is of rank one, $\mathbf{S}_W^{-1}\mathbf{S}_E$ is also of rank one. Thus, the eigenvalue decomposition of $\mathbf{S}_W^{-1}\mathbf{S}_E$ leads to only one non-zero eigenvalue, and hence FD can only reduce to dimension $d = 1$.

Loog and Duin have recently proposed a new LDR technique for normally distributed classes [12], namely LD, which takes the Chernoff distance in the original space into consideration to minimize the error rate in the transformed space. They consider the concept of *directed distance matrices*, and a linear transformation in the original space, to finally generalize Fisher's criterion in the transformed space by substituting the between-class scatter matrix for the corresponding directed distance matrix. The LD criterion consists of obtaining the matrix $\mathbf{A}$ that maximizes the function [12]:

$$J_{LD_2}(\mathbf{A}) = \mathrm{tr}\left\{(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1}\left[\mathbf{A}\mathbf{S}_E\mathbf{A}^t - \mathbf{A}\mathbf{S}_W^{1/2}\frac{p_1\log(\mathbf{S}_W^{-1/2}\mathbf{S}_1\mathbf{S}_W^{-1/2}) + p_2\log(\mathbf{S}_W^{-1/2}\mathbf{S}_2\mathbf{S}_W^{-1/2})}{p_1 p_2}\mathbf{S}_W^{1/2}\mathbf{A}^t\right]\right\}, \quad (3)$$

where the logarithm of a matrix $\mathbf{M}$, $\log(\mathbf{M})$, is defined as

$$\log(\mathbf{M}) \triangleq \mathbf{\Phi}\log(\mathbf{\Lambda})\mathbf{\Phi}^{-1}, \quad (4)$$

with $\mathbf{\Phi}$ and $\mathbf{\Lambda}$ representing the eigenvectors and eigenvalues of $\mathbf{M}$.

The solution to this criterion is given by the matrix $\mathbf{A}$ that is composed of the d eigenvectors (whose eigenvalues are the largest ones) of the following matrix:

$$\mathbf{S}_{LD_2} = \mathbf{S}_W^{-1}\left[\mathbf{S}_E - \mathbf{S}_W^{1/2}\frac{p_1\log(\mathbf{S}_W^{-1/2}\mathbf{S}_1\mathbf{S}_W^{-1/2}) + p_2\log(\mathbf{S}_W^{-1/2}\mathbf{S}_2\mathbf{S}_W^{-1/2})}{p_1 p_2}\mathbf{S}_W^{1/2}\right]. \quad (5)$$

The FD criterion discussed above aims to minimize the classification error by maximizing the Mahalanobis distance between distributions, resulting in an optimal criterion (in the Bayesian context) only when the covariance matrices are equal [13], while the LD criterion utilizes, as pointed out above, a directed distance matrix, which is incorporated in Fisher's criterion assuming the within-class scatter matrix is the identity.

1.2. Highlights of the proposed criterion

In this paper, we take advantage of the relationship between the probability of classification error of the optimal (in the Bayesian sense) classifier and the Chernoff distance, and propose a new criterion for LDR that aims to maximize the separability of the distributions in the transformed space based on the Chernoff measure. Since we are assuming the original distributions are normal, the distributions in the transformed space are also normal.[2] Thus, the Bayes classifier in the transformed space is quadratic and the classification error (also known as *true* error [1]) does not have a closed-form expression. Let $p(\mathbf{y}|\omega_i)$ be the class-conditional probability that a vector $\mathbf{y}=\mathbf{A}\mathbf{x}$ in the transformed space belongs to class ω_i. The probability of error can be bounded by the Chernoff distance between two distributions as follows [1]:

$$\Pr[\text{error}] = \int_{\mathscr{R}_2} p_1 p(\mathbf{y}|\omega_1)\,d\mathbf{y} + \int_{\mathscr{R}_1} p_2 p(\mathbf{y}|\omega_2)\,d\mathbf{y} \quad (6)$$

$$\leqslant p_1^{\beta} p_2^{1-\beta}\int p^{\beta}(\mathbf{y}|\omega_1)p^{1-\beta}(\mathbf{y}|\omega_2)\,d\mathbf{y} = p_1^{\beta} p_2^{1-\beta} e^{-k(\beta,\mathbf{A})}, \quad (7)$$

where $\mathscr{R}_1$ and $\mathscr{R}_2$ are the regions in which an object is assigned to class ω_1 or ω_2, respectively. For normally distributed classes, it can be shown that the Chernoff distance is given by [1]:

$$k(\beta,\mathbf{A}) = \frac{\beta(1-\beta)}{2}(\mathbf{A}\mathbf{m_1} - \mathbf{A}\mathbf{m_2})^t[\beta\mathbf{A}\mathbf{S}_1\mathbf{A}^t + (1-\beta)\mathbf{A}\mathbf{S}_2\mathbf{A}^t]^{-1}(\mathbf{A}\mathbf{m_1} - \mathbf{A}\mathbf{m_2}) + \frac{1}{2}\log\frac{|\beta\mathbf{A}\mathbf{S}_1\mathbf{A}^t + (1-\beta)\mathbf{A}\mathbf{S}_2\mathbf{A}^t|}{|\mathbf{A}\mathbf{S}_1\mathbf{A}^t|^{\beta}|\mathbf{A}\mathbf{S}_2\mathbf{A}^t|^{1-\beta}}, \quad (8)$$

where $\beta \in [0, 1]$.

The larger the value of $k(\beta, \mathbf{A})$ is, the smaller the bound for the classification error is, and hence, in this paper, we propose to maximize (8). To clarify this, we note that the FD criterion also aims to maximize the separability between distributions in the transformed space, but coincides with the optimal classifier only when the latter is linear, i.e. when the covariance matrices are coincident, a rare case. As observed above, the LD criterion utilizes the Chernoff distance in its directed distance matrix but in the original space. This criterion, however, does not optimize such a distance in the *transformed* space, as it can be observed in the example given below. A few remarks are discussed prior to the example.

For normally distributed classes, Eqs. (7) and (8) are useful for approximating the probability of error for the *optimal* (Bayesian) classifier. Since this is not usually the case for real-life data, other factors should be taken into consideration. First, normal distributions are characterized by the first two moments, while it is not (always) the case for real-life data. As pointed out in Ref. [1], the Chernoff bound can still be used when normality is not in place; however, it is not as accurate as for normal

[2] We note, however, that this assumption is not necessarily true in practice, and that our proposed criterion is still efficient even when data has other distributions, as shown in the empirical result section.