Available online at www.sciencedirect.com
ScienceDirect
Neurocomputing 71 (2007) 455–458
NEUROCOMPUTING
www.elsevier.com/locate/neucom

Letters

Kernel subclass discriminant analysis

Bo Chen*, Li Yuan, Hongwei Liu, Zheng Bao

National Lab of Radar Signal Processing, Xidian University, Xi'an, Shaanxi 710071, China

Received 3 February 2007; received in revised form 2 June 2007; accepted 23 July 2007
Communicated by L.C. Jain
Available online 2 September 2007

Abstract

In order to overcome the restricts of linear discriminant analysis (LDA), such as multivariate Normal distributed classes with equal covariance matrix but different means and the single-cluster structure in each class, subclass discriminant analysis (SDA) is proposed recently. In this paper the kernel SDA is presented, called KSDA. Moreover, we reformulate SDA so as to avoid the complicated derivation in the feature space. The encouraging experimental results on eight UCI data sets demonstrate the efficiency of our method.

Keywords: Linear discriminant analysis (LDA); Kernel linear discriminant analysis (KLDA); Subclass discriminant analysis (SDA); Feature space; Kernel methods

1. Introduction

Linear discriminant analysis (LDA) is a popular method for linear dimensionality reduction, which maximizes between-class scatter and minimizes within-class scatter. However, LDA is optimal only in the case that all the classes are generated from underlying multivariate Normal distributions of common covariance matrix but different means and each class is expressed by a single cluster [4,12], therefore LDA cannot give desired results for multimodally distributed data sets, such as face recognition [5], radar automatic target recognition [3] and so on. In order to overcome the limitations of LDA, recently Zhu and Martinez [12] propose subclass discriminant analysis (SDA). Just as its name implies, SDA divides each class into suitable subclasses so as to approximate the underlying distribution with mixture of Gaussians and then performs LDA among these subclasses, moreover, the authors also employ a stability criteria [8] to determine the optimal subclass divisions. In this letter we develop SDA into kernel SDA (KSDA) in the feature space, which can result in a better subspace for the classification task since a nonlinear clustering technique can find the underlying subclasses more exactly in the feature space and nonlinear LDA [5] can provide a nonlinear discriminant hyperplane. Furthermore, a reformulation of SDA is given to avoid the complicated derivation in the feature space.

*Corresponding author. Tel.: +86 29 88209504.
E-mail address: bchen@mail.xidian.edu.cn (B. Chen).

2. Subclass discriminant analysis

In SDA, a new between-subclass scatter matrix is defined

$$S_{\mathrm{SDA}}^{(b)} = \sum_{i=1}^{C-1}\sum_{j=1}^{H_i}\sum_{k=i+1}^{C}\sum_{l=1}^{H_k} p_{ij}p_{kl}(\mu_{ij}-\mu_{kl})(\mu_{ij}-\mu_{kl})^{\mathrm{T}}, \quad (1)$$

where C is the number of data classes, H_i the number of subclass divisions in class i, n the total number of all samples, n_{ij} the number of the jth subclass in class i, $p_{ij} = n_{ij}/n$ and μ_{ij} are the prior and mean of the jth subclass in class i.

In order to determine the optimal subclass divisions, the authors propose a discriminant stability criteria [12,8], which can evaluate whether LDA works

$$G = \sum_{i=1}^{m}\sum_{j=1}^{i}(u_j^{\mathrm{T}} w_i)^2, \quad (2)$$

where w_i is the i eigenvector of between-subclass scatter matrix $S_{\mathrm{SDA}}^{(b)}$ and u_j the jth eigenvector of the covariance matrix of the data [12] (which is defined as

0925-2312/$ - see front matter
doi:10.1016/j.neucom.2007.07.006

$S^{(m)}_{\mathrm{SDA}} = \sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^{\mathrm{T}}$, μ is the global mean), $m < rank(S^{(b)}_{\mathrm{SDA}})$. The smaller the value of cost function (2), the better SDA can work with the current subclass divisions.

Since SDA is just LDA when $H_i = 1$ for $i = 1, 2, \ldots, C$, LDA can be regarded as a special case of SDA. From (1), we know that $rank(S^{(b)}_{\mathrm{SDA}}) \leqslant \min(H-1, p)$, where $H = \sum_{i=1}^{C} H_i$ is the total number of subclasses and p the dimensionality of the range of the covariance matrix. Therefore, SDA can also solve the problem posed by the deficiency of the rank of the ordinary between-class scatter matrix, which has been proved by the example in [12].

3. Kernel SDA

In this section, we present a nonlinear extension of SDA based on kernel functions, KSDA. The main idea of the kernel method is that without knowing the nonlinear feature mapping explicitly, we can work on the feature space through kernel functions, as long as the problem formulation depends only on the inner products between data points. This is based on the fact that for any kernel function κ satisfying Mercer's condition [3] there exists a mapping Φ such that $\langle \Phi(x), \Phi(y) \rangle = \kappa(x, y)$, where $\langle , \rangle$ is an inner product in the feature space F transformed by Φ.

We apply the kernel method to perform SDA in the feature space instead of the original input space. Given a kernel function κ, let Φ be a mapping satisfying (3). So $S^{(b)}_{\mathrm{KSDA}}$ and $S^{(m)}_{\mathrm{KSDA}}$ in the feature space F can be expressed as

$$S^{(b)}_{\mathrm{KSDA}} = \sum_{i=1}^{C-1}\sum_{j=1}^{H_i}\sum_{k=i+1}^{C}\sum_{l=1}^{H_k} p_{ij}p_{kl}(\overline{\Phi}_{ij} - \overline{\Phi}_{kl})(\overline{\Phi}_{ij} - \overline{\Phi}_{kl})^{\mathrm{T}}, \quad (3)$$

$$S^{(m)}_{\mathrm{KSDA}} = \sum_{i=1}^{n}(\Phi_i - \overline{\Phi})(\Phi_i - \overline{\Phi})^{\mathrm{T}}, \quad (4)$$

where $\overline{\Phi}_{ij}$ indicates the mean vector of jth subclass of ith class, $\overline{\Phi}$ is the global mean.

Similar to SDA, KSDA maximizes $|\mathbf{V}^{\mathrm{T}} S^{(b)}_{\mathrm{KSDA}} \mathbf{V}| / |\mathbf{V}^{\mathrm{T}} S^{(m)}_{\mathrm{KSDA}} \mathbf{V}|$ to find a transformation matrix $\mathbf{V}$, the columns of which are the eigenvectors corresponding to the $r \leqslant \min(H-1, p)$ largest eigenvalues of

$$S^{(b)}_{\mathrm{KSDA}}\mathbf{V} = \lambda S^{(m)}_{\mathrm{KSDA}}\mathbf{V}. \quad (5)$$

Let the transformation matrix $\mathbf{V}$ be represented as

$$\mathbf{V} = \mathbf{\Phi}\boldsymbol{\alpha}, \quad (6)$$

where $\mathbf{\Phi} = [\Phi_1, \ldots, \Phi_n]$, $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_r]$.

Usually one substitutes (3), (4), and (6) into (5) to obtain the equation represented by the kernel Gram matrix; however, the whole derivation procedures and representations of kernel scatter matrices are complicated and not intuitive [1]. Therefore, in order to simplify it we will give a new representation of SDA based on the following scatter matrices

$$S^{(b)}_{\mathrm{SDA}} = \mathbf{X}\mathbf{D}^{(b)}_{\mathrm{SDA}}\mathbf{X}^{\mathrm{T}}, \quad (7)$$

$$S^{(m)}_{\mathrm{SDA}} = \mathbf{X}\mathbf{D}^{(m)}_{\mathrm{SDA}}\mathbf{X}^{\mathrm{T}}, \quad (8)$$

where

$$\mathbf{D}^{(b)}_{\mathrm{SDA}(i,j)} = \begin{cases} (n-n_k)/(n^2 \cdot n_{kl}) & \text{if } z_i = z_j = C_{kl} \\ 0 & \text{if } z_i \neq z_j \text{ but } y_i = y_j, \\ -1/n^2 & \text{if } y_i \neq y_j, \end{cases} \quad (9)$$

$$\mathbf{D}^{(m)}_{\mathrm{SDA}(i,j)} = \begin{cases} 1 - 1/n & \text{if } x_i = x_j, \\ -1/n & \text{else}, \end{cases} \quad (10)$$

where $\mathbf{X} = [x_1, \ldots, x_n]$, $y_i \in [1, C]$ is the class label of the sample x_i, $n_k = \sum_{l=1}^{H_k} n_{kl}$, C_{kl} indicates the lth subclass of kth class, z_i denotes the subclass which x_i belongs to.

If SDA is transformed into the feature space, (7) and (8) can be modified as

$$S^{(b)}_{\mathrm{KSDA}} = \mathbf{\Phi}\mathbf{D}^{(b)}_{\mathrm{KSDA}}\mathbf{\Phi}^{\mathrm{T}}, \quad (11)$$

$$S^{(m)}_{\mathrm{KSDA}} = \mathbf{\Phi}\mathbf{D}^{(m)}_{\mathrm{KSDA}}\mathbf{\Phi}^{\mathrm{T}}. \quad (12)$$

Substituting (6), (11), and (12), into (5), we obtain

$$\mathbf{\Phi}\mathbf{D}^{(b)}_{\mathrm{KSDA}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\boldsymbol{\alpha} = \lambda\mathbf{\Phi}\mathbf{D}^{(m)}_{\mathrm{KSDA}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\boldsymbol{\alpha}. \quad (13)$$

Then multiplying (13) by $\mathbf{\Phi}^{\mathrm{T}}$ from the left-hand side, we have

$$\mathbf{K}\mathbf{D}^{(b)}_{\mathrm{KSDA}}\mathbf{K}\boldsymbol{\alpha} = \lambda\mathbf{K}\mathbf{D}^{(m)}_{\mathrm{KSDA}}\mathbf{K}\boldsymbol{\alpha}, \quad (14)$$

where $\mathbf{K} \in \mathrm{R}^{n\times n}$ is the kernel Gram matrix. Compared with traditional forms of GDA [6], the new reformulation of KSDA is more concise and easy to operate. Note that $\mathbf{D}^{(b)}_{\mathrm{KSDA}}$ and $\mathbf{D}^{(m)}_{\mathrm{KSDA}}$ should reflect the relations of samples in the feature space, therefore, the division of subclasses has to be obtained by kernel clustering techniques. In this paper we use the kernel k-means method [10]. In addition, the stable criterion should also be reformulated in the feature space. Although the kernel scatter matrices may be of unknown dimensionalities, we can calculate the eigenvectors corresponding to the first largest eigenvalues of them using the kernel Gram matrix. The details can be referred to [9].

4. Experimental results

We demonstrate that our proposed method KSDA is an effective nonlinear extension of SDA by comparing the classification performances of KSDA and other kernel-based nonlinear discriminant analysis algorithms as well as GDA [1], Kernel direct LDA (KDDA) [7], complete kernel Fisher discriminant analysis (CKFD) [11] and kernel principal component analysis (KPCA) [10]. Eight data sets collected from UCI machine learning repository [2] and the ETH-80 database [6] were used, all of which were centered and normalized to a distribution with zero mean and unit