Neurocomputing 74 (2011) 568–574

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Generalized nonlinear discriminant analysis and its small sample size problems

Li Zhang [a,b,*], Wei Da Zhou [b], Pei-Chann Chang [c]

[a] Research Center of Machine Learning and Data Analysis, School of Computer Science and Technology, Soochow University, Suzhou 215006, Jiangsu, China
[b] Institute of Intelligent Information Processing, Xidian University, Xi'an 710071, Shaanxi, China
[c] Department of Information Management, Yuan Ze University, Taoyuan 32026, Taiwan, China

ARTICLE INFO

Article history:
Received 25 May 2009
Received in revised form
10 September 2010
Accepted 14 September 2010
Communicated by Liang Wang
Available online 27 October 2010

Keywords:
Fisher discriminant analysis
Kernel trick
Small sample size problem

ABSTRACT

This paper develops a generalized nonlinear discriminant analysis (GNDA) method and deals with its small sample size (SSS) problems. GNDA is a nonlinear extension of linear discriminant analysis (LDA), while kernel Fisher discriminant analysis (KFDA) can be regarded as a special case of GNDA. In LDA, an under sample problem or a small sample size problem occurs when the sample size is less than the sample dimensionality, which will result in the singularity of the within-class scatter matrix. Due to a high-dimensional nonlinear mapping in GNDA, small sample size problems arise rather frequently. To tackle this issue, this research presents five different schemes for GNDA to solve the SSS problems. Experimental results on real-world data sets show that these schemes for GNDA are very effective in tackling small sample size problems.

1. Introduction

Discriminant analysis has been widely used for feature extraction and dimensionality reduction in pattern recognition. Linear discriminant analysis (LDA), also known as Fisher linear discriminant is one of the most commonly used method [1]. The goal of LDA is to find an optimal subspace such that the separability of two classes is maximized. LDA is to maximize

$$\frac{\mathrm{tr}(\mathbf{W}^T\mathbf{S}_b\mathbf{W})}{\mathrm{tr}(\mathbf{W}^T\mathbf{S}_w\mathbf{W})} \tag{1}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of matrix $\cdot$, $\mathbf{W}$ is a linear projection or transformation matrix, $\mathbf{S}_b$ is the between-class scatter matrix and $\mathbf{S}_w$ is the within-class scatter matrix. Maximizing (1) results in the following generalized eigenvalue problem

$$\mathbf{S}_b\mathbf{W} = \lambda\mathbf{S}_w\mathbf{W} \tag{2}$$

The optimal discriminant subspace is spanned by the generalized eigenvectors. If $\mathbf{S}_w$ is nonsingular, the solution to the generalized eigenvalue problem (1) is obtained by applying eigendecomposition on $\mathbf{S}_w^{-1}\mathbf{S}_b$. However, for a small sample size (SSS) problem the scatter matrix $\mathbf{S}_w$ is singular. For example, face recognition is a kind of SSS problems with high-dimensional and few training samples. So far, there are some methods proposed to deal with the problem of singularity of $\mathbf{S}_w$, such as Fisherface [2], discriminant common vectors [3], dual space [4], LDA-GSVD (generalized singular value decomposition) [5], LDA-QR [6], PCA+NULL [7], and LDA-FKT (Fukunaga–Koontz transform) [8]. In [8], a unifying framework is proposed to understand different methods. By using Fukunaga–Koontz transform (FKT), the whole sample space can be decomposed into four subspaces. Discriminant information in the four subspaces is different, and the performance of methods depends on their subspaces. The authors in [8] also report that LDA-GSVD is equivalent to the LDA-FKT, and LDA-FKT has the best performance.

Unfortunately, LDA can only extract linear features from samples, it fails to process the data which consist of nonlinear features [9]. Kernel Fisher discriminant analysis (KFDA), one of the nonlinear discriminant methods, has been developed for extracting nonlinear discriminant features [9]. A similar work as KFDA is presented in [10]. Kernel functions are restricted to positive semi-definite symmetric functions, i.e., Mercer kernels as in [11,12]. KFDA often encounters SSS problems because $\mathbf{S}_w$ in a high-dimensional feature space is always singular. To overcome the computational difficulty with KFDA, a perturbation $\mu\mathbf{I}$ is added to $\mathbf{S}_w$ in [9] where $\mathbf{I}$ is an identity matrix as the same size as $\mathbf{S}_w$, and kernel Fisherface is proposed in [13].

This paper proposes a generalized nonlinear discriminant analysis (GNDA). GNDA consists of two steps. First, data in a sample space are mapped into a nonlinear mapping space by using some nonlinear mapping function. Then LDA is implemented in the nonlinear mapping space. In GNDA, the nonlinear mapping function can be any real-valued nonlinear function, for instance, empirical mapping functions as in [14,15], Mercer kernel mapping as in [11,12], etc. GNDA is identical

* Corresponding author at: Research Center of Machine Learning and Data Analysis, School of Computer Science and Technology, Soochow University, Suzhou 215006, Jiangsu, China.
E-mail address: zhangliml@suda.edu.cn (L. Zhang).

0925-2312/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2010.09.022

with KFDA if the same Mercer kernel is used in both of them. In this sense, GNDA suggests a universe framework which unifies KFDA. Similarly, SSS problems will occur when using GNDA. We solve this problem by extending linear methods in [2,5–8] to obtain GNDA-Fisherface, GNDA-PCA+NULL, GNDA-QR, GNDA-GSVD and GNDA-FKT algorithms.

The rest of the paper is organized as follows: Section 2 briefly reviews the related work including LDA and KFDA. Section 3 describes GNDA and gives a theorem to show that GNDA unifies KFDA under certain conditions. In Section 4, we cope with SSS problems in GNDA and present five schemes. Section 5 shows some simulation results of real-world recognition problems on the face and digit databases. The concluding remarks are given in Section 6.

2. Related work

In this section, we briefly review LDA and KFDA, respectively.

2.1. Linear discriminant analysis

Here we consider a two-class problem; $X_1=\{\mathbf{x}_1,\ldots,\mathbf{x}_{\ell_1}\}$ and $X_2=\{\mathbf{x}_1,\ldots,\mathbf{x}_{\ell_2}\}$ are two-class sample sets, respectively. Let $X=X_1\cup X_2$. The goal of LDA is to find an optimal linear transformation such that the separability of two classes is maximized. This is achieved by minimizing the within-class distance whilst maximizing the between-class distance. In Section 1, we have introduced the optimization formula for LDA (1). The between- and within-class scatter matrices can be written as

$$\mathbf{S}_b=(\mathbf{m}_1-\mathbf{m}_2)(\mathbf{m}_1-\mathbf{m}_2)^T \tag{3}$$

and

$$\mathbf{S}_w=\sum_{i=1}^{2}\sum_{j=1,\mathbf{x}_j\in X_i}^{l_i}(\mathbf{x}_j-\mathbf{m}_i)(\mathbf{x}_j-\mathbf{m}_i)^T \tag{4}$$

where $\mathbf{m}_i=(1/\ell_i)\sum_{j=1,\mathbf{x}_j\in X_i}^{\ell_i}\mathbf{x}_j, i=1,2$ is the mean of the i-th class samples. Maximizing (1) results in the generalized eigenvalue problem (2). The above LDA method for two-class problem can be directly generalized to multi-class problem [1].

2.2. Kernel Fisher discriminant analysis

KFDA is one of the nonlinear discriminant analysis methods [9]. For a two-class problem, KFDA has the following form:

$$\max_{\boldsymbol{\alpha}}\frac{\mathrm{tr}(\boldsymbol{\alpha}^T\mathbf{S}_b^K\boldsymbol{\alpha})}{\mathrm{tr}(\boldsymbol{\alpha}^T\mathbf{S}_w^K\boldsymbol{\alpha})} \tag{5}$$

where $\mathbf{S}_w^K$ and $\mathbf{S}_b^K$ are quasi within-class and between-class scatter matrices, respectively. There have

$$\mathbf{S}_w^K=\mathbf{KMK} \tag{6}$$

and

$$\mathbf{S}_b^K=\sum_{m=1}^{2}\ell_i(\mathbf{K}_m\mathbf{e}_m-\mathbf{Ke})(\mathbf{K}_m\mathbf{e}_m-\mathbf{Ke})^T \tag{7}$$

where $\mathbf{K}$ and $\mathbf{K}_m$ are kernel gram matrices, $(\mathbf{K})_{ij}=k(\mathbf{x}_i,\mathbf{x}_j)|_{\mathbf{x}_i,\mathbf{x}_j\in X}$, $(\mathbf{K}_m)_{ij}=k(\mathbf{x}_i,\mathbf{x}_j)|_{\mathbf{x}_i\in X,\mathbf{x}_j\in X_m}$, $\mathbf{e}_m=[1/\ell_m,\ldots,1/\ell_m]^T\in\mathbb{R}^{\ell_m\times 1}$, $\mathbf{e}=[1/\ell,\ldots,1/\ell]^T\in\mathbb{R}^{\ell\times 1}$, $\mathbf{M}=\mathbf{I}-\mathbf{N}$, $\mathbf{I}\in\mathbb{R}^{\ell\times\ell}$ is an identify matrix, and

$$\mathbf{N}_{ij}=\begin{cases}1/\ell_m & \text{if } \mathbf{x}_i,\mathbf{x}_j\in X_m\\ 0 & \text{otherwise}\end{cases}$$

It can be shown that $\mathbf{MM}=\mathbf{M}$ and $\mathbf{M}^T=\mathbf{M}$. In other words, $\mathbf{M}$ is an idempotent matrix. Similarly, maximizing (5) results in the following generalized eigenvalue problem

$$\mathbf{S}_b^K\boldsymbol{\alpha}=\gamma\mathbf{S}_w^K\boldsymbol{\alpha} \tag{8}$$

KFDA can also be extended to solve multi-class problems.

3. Generalized nonlinear discriminant analysis

This section presents GNDA which implements a LDA in a nonlinear mapping space. The corresponding nonlinear mapping function $g(\mathbf{x})$ can be any real-valued nonlinear function. Let the set of independently and identically distributed samples be

$$\{(\mathbf{x}_i,y_i)|\mathbf{x}_i\in\mathbb{R}^n,\ y_i\in\{1,2,\ldots,C\},\ i=1,\ldots,\ell\}$$

where C is the total number of classes, ℓ is the total number of samples and n is the dimensionality of samples. Let X_m be the m-th class sample set; then the training set is $X=\bigcup_{m=1}^{C}X_m$. The number of samples in the m-th class is denoted by ℓ_m; thus $\ell=\sum_{m=1}^{C}\ell_m$. The set of mapped samples in the nonlinear mapping space can be expressed as

$$\{(g(\mathbf{x}_i),y_i)|g(\mathbf{x}_i)\in\mathbb{R}^N,\ y_i\in\{1,2,\ldots,C\},\ i=1,\ldots,\ell\} \tag{9}$$

where $g(\mathbf{x})$ is a pre-specified real nonlinear mapping, and N is the dimensionality of the nonlinear mapping space. Let the sample matrix in the nonlinear mapping space be

$$\mathbf{G}=[g(\mathbf{x}_1),g(\mathbf{x}_2),\ldots,g(\mathbf{x}_\ell)]\in\mathbb{R}^{N\times\ell} \tag{10}$$

Since the mapped patterns in the nonlinear mapping space are definitely known when the training samples are given. The computation of any statistic about the samples in the nonlinear mapping space is feasible, such as the mean of samples, which is impossible in a reproducing kernel Hilbert space (RKHS). In the nonlinear mapping space, the Fisher criterion is to maximize

$$J_G(\mathbf{W})=\frac{\mathrm{tr}(\mathbf{W}^T\mathbf{S}_b^G\mathbf{W})}{\mathrm{tr}(\mathbf{W}^T\mathbf{S}_w^G\mathbf{W})} \tag{11}$$

where the between-class scatter matrix is

$$\mathbf{S}_b^G=\sum_{m=1}^{C}\ell_m(\mathbf{m}_m-\mathbf{m})(\mathbf{m}_m-\mathbf{m})^T=\sum_{m=1}^{C}\ell_m(\mathbf{G}_m\mathbf{e}_m-\mathbf{Ge})(\mathbf{G}_m\mathbf{e}_m-\mathbf{Ge})^T \tag{12}$$

in which $\mathbf{G}_m$ is the m-th class sample matrix consisting of column vector $g(\mathbf{x}_i)|_{\mathbf{x}_i\in X_m}$, and the within-class scatter matrix is

$$\mathbf{S}_w^G=\sum_{m=1}^{C}\sum_{i=1,\mathbf{x}_i\in X_m}^{\ell_m}(g(\mathbf{x}_i)-\mathbf{m}_m)(g(\mathbf{x}_i)-\mathbf{m}_m)^T=\mathbf{GMG}^T \tag{13}$$

Hence the total scatter matrix in the nonlinear mapping space can be expressed as

$$\mathbf{S}_t^G=\mathbf{S}_b^G+\mathbf{S}_w^G \tag{14}$$

If $\mathbf{S}_w^G$ is nonsingular, the solution to (11) amounts to solving the generalized eigenvalue problem

$$\mathbf{S}_b^G\mathbf{W}=\lambda\mathbf{S}_w^G\mathbf{W} \tag{15}$$

Let the eigenvalues of (15) be $\{\lambda_i\}$ and the corresponding eigenvectors be $\{\boldsymbol{v}_i\}$. Sort eigenvectors $\{\boldsymbol{v}_i\}$ according to λ_i in descending order $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_{C-1}\geq\cdots$. Let $\mathbf{W}$ consist of the first $C-1$ eigenvectors. Namely there have

$$\mathbf{W}=[\boldsymbol{v}_1,\ldots,\boldsymbol{v}_{C-1}]\in\mathbb{R}^{N\times(C-1)} \tag{16}$$

Specifically, if the nonlinear mapping function takes an empirical mapping function, then the nonlinear mapping space is an empirical mapping space or a hidden space [14,15]. Of course, kernel functions can be used to construct nonlinear mapping functions and are not constrained to Mercer's condition. It is known that