Signal Processing 88 (2008) 2293–2301

Contents lists available at ScienceDirect

Signal Processing

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

Solution for supervised graph embedding: A case study

Qubo You *, Nanning Zheng, Ling Gao, Shaoyi Du, Jianyi Liu

Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, Shaanxi Province 710049, PR China

ARTICLE INFO

Article history:
Received 13 July 2007
Received in revised form
21 February 2008
Accepted 19 March 2008
Available online 27 March 2008

Keywords:
Graph embedding
Discriminant transformation
Null space
Range space
SSS problem

ABSTRACT

Recently, Graph Embedding Framework has been proposed for feature extraction. Although many algorithms can be used to extract the discriminant transformation of supervised graph embedding, it is still an open issue which algorithm is more robust. In this paper, we first review the classical algorithms which can extract the discriminant transformation of linear discriminant analysis (LDA), and then generalize these classical algorithms for computing the discriminant transformation of supervised graph embedding. Secondly, we theoretically analyze the robustness of these generalized algorithms. Finally, in order to overcome the disadvantages of these generalized algorithms, we propose an effective method, called total space solution for supervised graph embedding (TSS/SGE), to extract the robust discriminant transformation of Supervised Graph Embedding. Extensive experiments and comprehensive comparison on real-world data are performed to demonstrate the robustness of our proposed TSS/SGE.

1. Introduction

An important issue in the field of pattern recognition is feature extraction and dimensionality reduction. A common way to resolve this problem is discriminant subspace analysis without losing a lot of significant information [1,2]. In this field, linear models have played an important role because of their simplification and analytical tractability [3]. In the past decades, many algorithms have been proposed. Among these algorithms, linear discriminant analysis (LDA) [4] is one of the most popular techniques. Furthermore, a large family of algorithms, such as local discriminant embedding (LDE) [5] and marginal Fisher analysis (MFA) [6], has been developed based on manifold learning [7]. Although motivations of the above algorithms are different, their final objectives are similar. In 2007, Yan et al. [6] presented a general framework (Graph Embedding) to unify the above algorithms for dimensionality reduction. In this framework, the final discriminant transformation $\mathbf{Q}^*$ can be computed from the following generalized eigenvalue problem [6]:

$$\mathbf{M}_b\mathbf{q} = \lambda\mathbf{M}_w\mathbf{q} \tag{1}$$

where $\mathbf{M}_b$ and $\mathbf{M}_w$ define the measure to be maximized and that to be minimized, respectively. In this framework, LDA is a special case, where $\mathbf{M}_b$ and $\mathbf{M}_w$ are the between-class scatter matrix $\mathbf{S}_b$ and the within-class scatter matrix $\mathbf{S}_w$, respectively. Since $\mathbf{M}_w$ is singular in the case of the SSS problem, $\mathbf{Q}^*$ cannot be directly computed by eigen-decompositions of $\mathbf{M}_w^{-1}\mathbf{M}_b$. Unfortunately, the SSS problem frequently occurs in many applications, such as information retrieval and face recognition [4–6].

In the literature of LDA, different variations of LDA have been proposed to deal with the SSS problem. In general, the discriminant transformation of LDA can be extracted from pseudo-inverse [8], generalized singular value decomposition (GSVD) [9–12], the range space of $\mathbf{S}_w$ [3], the null space of $\mathbf{S}_w$ [13], the range space of $\mathbf{S}_b$ [14], the range space of $\mathbf{S}_t$ [11,15,16] or the dual-space of $\mathbf{S}_w$ [17]. Since LDA is a special case in the framework of supervised graph embedding (SGE), most of algorithms, extracting the discriminant transformation of LDA, can be generalized into this general framework. We can use these

* Corresponding author. Tel.: +86 29 82668802x8002; fax: +86 29 82668672.
E-mail address: youqubo@gmail.com (Q. You).

0165-1684/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2008.03.014

generalized algorithms to extract the discriminant transformation of specific SGE algorithms, such as LDE [5] and MFA [6].

However, up to the present, it is still an open issue which algorithm is more robust. Is it from pseudo-inverse, GSVD, the range space of $\mathbf{M}_w$, the null space of $\mathbf{M}_w$, the range space of $\mathbf{M}_b$, the range space of $\mathbf{M}_t = \mathbf{M}_w + \mathbf{M}_b$, or the dual-space of $\mathbf{M}_w$? Using the range space of $\mathbf{M}_w$, the algorithms do not use the discriminative information of the null space; while using the null space of $\mathbf{M}_w$, the algorithms lose the important discriminative information of the range space. Similarly, the algorithms, using the range space of $\mathbf{M}_b$ or the dual-space of $\mathbf{M}_w$, also result in the loss of important discriminative information.

In this paper, we focus on the issue which algorithm is appropriate for extracting the discriminant transformation in the framework of SGE. Firstly, we review the classical algorithms which can extract the discriminant transformation of LDA. Secondly, we generalize these classical algorithms for computing the discriminant transformation of SGE, and then theoretically analyze the robustness of these generalized algorithms. Finally, in order to overcome the disadvantages of these generalized algorithms, we propose an effective method, called total space solution for supervised graph embedding (TSS/SGE), to extract the robust discriminant transformation of SGE. Different from those generalized algorithms, our proposed TSS/SGE uses all the information in the total space of $\mathbf{M}_w$ to extract the discriminant transformation. Therefore, the discriminant transformation of TSS/SGE can contain more important information than those generalized algorithms. Extensive experiments and comprehensive comparison on real-world data are performed to demonstrate the robustness of our proposed TSS/SGE.

The rest of the paper is organized as follows: classical algorithms are briefly reviewed in Section 2. In Section 3, we generalize these classical algorithms for computing the discriminant transformation of SGE, and then theoretically analyze the robustness of these generalized algorithms. TSS/SGE is described in Section 4. In Section 5, extensive experiments and comprehensive comparison are performed to demonstrate the robustness of our proposed TSS/SGE. Finally, conclusions are summarized in Section 6.

2. An overview of classical algorithms

In this section, we give a brief overview of classical algorithms which can extract the discriminant transformation of LDA. These classical algorithms include PCA + LDA [4], pseudo Fisher linear discriminant analysis (PFLDA) [8], LDA/GSVD [9], LDA/FKT [16], PCA + Null Space Method (PCA + NSM) [13] and Robust Algorithm (RA) [3]. For the convenience of understanding, in the following, the small *italic* letters denote scalars, such as a, b, c; the small **bold** non-italic letters denote vectors, such as $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$; and the capital **bold** non-italic letters denote matrices, such as $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$. Let we have n samples $\{\mathbf{x}_i|\mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^n$ belonging to c classes, the number of samples in the ith class is n_i, and $\mathbf{x}_j^i$ denotes the jth sample in the ith class.

Defining

$$\mathbf{H}_b = [\sqrt{n_1}(\mu_1 - \mu), \sqrt{n_2}(\mu_2 - \mu), \ldots, \sqrt{n_c}(\mu_c - \mu)] \in \mathbb{R}^{d \times c} \tag{2}$$

$$\mathbf{H}_w = [\mathbf{X}_1 - \mu_1 e_1^{\mathrm{T}}, \mathbf{X}_2 - \mu_2 e_2^{\mathrm{T}}, \ldots, \mathbf{X}_c - \mu_c e_c^{\mathrm{T}}] \in \mathbb{R}^{d \times n} \tag{3}$$

$$\mathbf{H}_t = [\mathbf{x}_1 - \mu, \mathbf{x}_2 - \mu, \ldots, \mathbf{x}_n - \mu] \in \mathbb{R}^{d \times n} \tag{4}$$

where μ is the mean vector of all the samples, μ_i is the mean vector of the samples in the ith class, $\mathbf{X}_i$ is the data matrix in the ith class, and $e_i = [1, 1, \ldots, 1]^{\mathrm{T}} \in^{n_i \times 1}$. Because of $\mathbf{S}_w = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (\mathbf{x}_j^i - \mu_i)(\mathbf{x}_j^i - \mu_i)^{\mathrm{T}}$, $\mathbf{S}_b = \sum_{i=1}^{c} n_i(\mu_i - \mu)(\mu_i - \mu)^{\mathrm{T}}$ and $\mathbf{S}_t = \mathbf{S}_w + \mathbf{S}_b$, we can find that $\mathbf{S}_b = \mathbf{H}_b\mathbf{H}_b^{\mathrm{T}}$, $\mathbf{S}_w = \mathbf{H}_w\mathbf{H}_w^{\mathrm{T}}$ and $\mathbf{S}_t = \mathbf{H}_t\mathbf{H}_t^{\mathrm{T}}$.

2.1. PCA + LDA

The common way, dealing with the SSS problem in LDA, is PCA + LDA [4], which first uses PCA to reduce the dimension of the original data before using classical LDA. The reduced dimension for PCA is chosen to make the within-class scatter matrix $\tilde{\mathbf{S}}_w$ nonsingular. In PCA + LDA, however, it is difficult to choose the reduced dimension for PCA. Furthermore, the PCA stage in PCA + LDA may result in the loss of important discriminative information.

2.2. PFLDA

Based on the pseudo-inverse of the within-class scatter matrix, PFLDA was proposed [8]. In PFLDA, the discriminant transformation $\mathbf{Q}^*$ can be extracted by eigen-decomposition of $\mathbf{S}_w^+\mathbf{S}_b$, where $\mathbf{S}_w^+$ is the pseudo-inverse of $\mathbf{S}_w$. Let $\mathbf{S}_w = \mathbf{U}_w\mathbf{\Sigma}_w\mathbf{U}_w^{\mathrm{T}}$, where $\mathbf{\Sigma}_w = \mathrm{diag}(\lambda_1^w, \ldots, \lambda_r^w)$, $\mathbf{U}_w = [\mathbf{u}_1^w, \ldots, \mathbf{u}_r^w]$, λ_i^w and $\mathbf{u}_i^w$ are the ith largest eigenvalue and the corresponding eigenvector of $\mathbf{S}_w$, and $r = \mathrm{rank}(\mathbf{S}_w)$. The pseudo-inverse of $\mathbf{S}_w$ can be computed as $\mathbf{S}_w^+ = \mathbf{U}_w\mathbf{\Sigma}_w^{-1}\mathbf{U}_w^{\mathrm{T}}$ [2].

2.3. PCA + NSM

In 1997, Belhumeur et al. [4] first briefly mentioned the idea of the NSM by the following objective function:

$$\mathbf{Q}^* = \arg \max_{\mathrm{trace}(\mathbf{Q}^{\mathrm{T}}\mathbf{S}_w\mathbf{Q})=0} \frac{\mathrm{trace}(\mathbf{Q}^{\mathrm{T}}\mathbf{S}_b\mathbf{Q})}{\mathrm{trace}(\mathbf{Q}^{\mathrm{T}}\mathbf{Q})} \tag{5}$$

In this idea, the samples are first projected onto the null space of $\mathbf{S}_w$. Then the transformation is computed by maximizing the between-class scatter of projected samples. Different from the NSM, PCA + NSM [13] was proposed to improve the efficiency of the NSM. Since the null space of $\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w$ is the intersection of the null space of $\mathbf{S}_b$ and the null space of $\mathbf{S}_w$, the samples are first projected onto the range space of $\mathbf{S}_t$ without loss of any discriminative information. Then the transformation is extracted in the new low dimensional space by the NSM.