Pattern Recognition 45 (2012) 1471–1481

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

# A novel supervised dimensionality reduction algorithm: Graph-based Fisher analysis

Yan Cui [a,b], Liya Fan [b,*]

[a] School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China
[b] School of Mathematics Sciences, Liaocheng University, Liaocheng 252059, China

ARTICLE INFO

*Article history:*
Received 29 June 2010
Received in revised form
6 September 2011
Accepted 7 October 2011
Available online 10 November 2011

*Keywords:*
Dimensionality reduction
Intrinsic graph
Penalty graph
Positive definite kernels
Indefinite kernels

ABSTRACT

In this paper, a novel supervised dimensionality reduction (DR) algorithm called graph- based Fisher analysis (GbFA) is proposed. More specifically, we redefine the intrinsic and penalty graph and trade off the importance degrees of the same-class points to the intrinsic graph and the importance degrees of the not-same-class points to the penalty graph by a strictly monotone decreasing function; then the novel feature extraction criterion based on the intrinsic and penalty graph is applied. For the non-linearly separable problems, we study the kernel extensions of GbFA with respect to positive definite kernels and indefinite kernels, respectively. In addition, experiments are provided for analyzing and illustrating our results.



## 1. Introduction

Techniques for dimensionality reduction in supervised or unsupervised learning tasks have attracted much attention in computer vision and pattern recognition. Among them, principal component analysis (PCA) [1] and linear discriminant analysis (LDA) [2] are two most popular linear subspace learning methods. PCA is an unsupervised learning algorithm, which performs dimensionality reduction by projecting the original $m$-dimensional data on to the $l\,(l \ll m)$-dimensional linear subspace spanned by the leading eigenvectors of the data's covariance matrix. While LDA is a supervised learning algorithm, it searches the projection axes on which the not-same-class points are far from each other while requiring the same-class points to be close to each other. Therefore, LDA encodes discriminating information in a linearly separable space. However a drawback of LDA is that it cannot be applied when the scatter matrix $S_w$ or $S_t$ is singular due to the small sample size problems. In the past, many LDA extensions have been developed to deal with this problem, such as, Pseudo-inverse LDA (PLDA) [3], regular LDA (RLDA) [4], Penalized discriminant analysis (PDA) [5], LDA/GSVD [6], LDA/QR [7], orthogonal LDA (OLDA) [8], null space LDA (NLDA) [9], direct-LDA(D-LDA) [10], CLDA [11] and two-stage LDA [12]. Despite the LDA-based algorithms having many applications, their effectiveness is still limited because the number of the available projection directions is lower than the class number. Furthermore, the LDA-based algorithms are proposed based upon the data approximately obeying a Gaussian distribution, which cannot always be satisfied in the real-world applications.

Recently, a number of graph-based DR learning methods have been successfully applied and became important methodologies in machine learning and pattern recognition fields. Compared with PCA and LDA, the graph-based algorithms need not assume that the data obeys a Gaussian distribution, so they are more general for discriminant analysis. Some known graph-based algorithms are locality preserving projection (LPP) [13], local linear embedding (LLE) [14], local Fisher discriminant analysis (LFDA) [15], Laplacianfaces [16] and unsupervised discriminant projection (UDP) [17], Marginal Fisher analysis (MFA) [18], Linear discriminant projection (LDP) [19], Graph-optimized locality preserving projections (GoLPP) [20] and Sparsity preserving discriminant analysis (SPDA) [21].

Despite the success of the graph-based embedding dimensionality reduction techniques, they still suffer from such issues as:

(1) The current learning algorithms in Refs. [13–16,18,20,21] are all based on the characterization of 'locality'. The local quantity suffices for modeling a single manifold, but does not suffice for modeling multi-manifold cases;
(2) From the definition of the local and nonlocal scatter matrices in Ref. [17], it disguises a discriminant disposal on the data in question, that is, the samples in the local neighborhood are

* Corresponding author. Tel.: +86 635 8239927.
*E-mail address:* fanliya63@126.com (L. Fan).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.10.006

deemed the same-class ones, while the samples in the nonlocal region are deemed the not-same-class ones. However, the so-called same-class or not-same-class is not generally identical to the true class information. Some sketch examples can be found in Ref. [22], meanwhile, the same question exists in Refs. [13,14,20,21];

(3) The local neighborhood of each sample in computing the adjacency matrix should adaptively be determined, since the local structure of each sample usually is different and depends on the data distribution. Therefore, they should not artificially be determined by the same criterion for all samples as in Refs. [13,14,16–19]. Not only does this ignore the actual data distribution, but also brings the difficulty of parameter selection.

Inspired by ideas in Refs. [2,15,17,19–22], in this paper, we will develop a novel supervised DR algorithm, called the Graph-based Fisher Analysis (GbFA), to overcome the limitation of [2,15, 17–19]. In contrast to [1–19], GbFA has the following advantages: (1) there is no assumption that the data obeys a Gaussian distribution, thus it is more general for discriminant analysis; (2) The intrinsic graph based on the same-class samples and the penalty graph based on the not-same-class samples are redefined, so GbFA encodes the discriminating information; (3) It adaptively trades off the importance degrees of the same-class samples to the intrinsic graph and the importance degrees of the not-same-class samples to the penalty graph by a strictly monotone decreasing function which encodes the distributing information of the samples as well; (4) Maximizing the criterion function of GbFA will make the original neighbor same-class samples much closer in the output space while pushing apart the original neighbor not-same-class samples in the output space, thus GbFA criterion will enhance the classification result of the data set. Moreover, based on the kernel trick in Refs. [23–26], we will extend the kernel GbFA model with positive definite kernels and indefinite kernels for the non-linearly separable problems, respectively. Furthermore, in contrast to Refs. [18,20,21], we redefine the adjacency matrix in the feature space, so it will be more reasonable in the real-world applications.

The organization of this paper is as follows. We present the GbFA criterion in Section 2. We discuss the connections between GbFA and other graph embedding DR algorithms in Section 3. We introduce the kernel GbFA with respect to positive definite kernels and indefinite kernels in Section 4. In Section 5, experiments are presented to demonstrate the effectiveness of GbFA and the kernel GbFA. Conclusions are summarized in Section 6.

## 2. Graph-based Fisher analysis

In this section, we introduce the basic ideas of GbFA and its algorithm derivation under the supervised scenarios. Supervised dimensionality reduction approaches aim to map the original data space to a lower dimensional space and preserve the class discriminatory information from the class labels.

### 2.1. Basic ideas

In order to improve the classification result, we want to construct an embedding map $V=[v_1,\ldots,v_l]\in\Re^{m\times l}$ such that the distance between the same-class samples is reduced in the output space while simultaneously the neighbor not-same-class samples are pushed apart to avoid the large overlaps of neighbor classes in the output space. For the given pattern samples data set $T=\{(x_1,y_1),\ldots,(x_n,y_n)\}\in X\times Y, X=[x_1,\ldots,x_n]\in\Re^{m\times n}$ is the input data matrix (or vertices set) and $Y=\{C_1,\ldots,C_c\}$ is the class label set. According to spectral graph theory [27], we attempt to construct an intrinsic graph $G=\{X,W\}$ and a penalty graph $G'=\{X,W'\}$, where $W\in\Re^{n\times n}$ and $W'\in\Re^{n\times n}$ are the adjacency matrices. With the embedding map $V$, the intra-class compactness can be characterized from the intrinsic graph by the term

$$G_c=\sum_{i=1}^{n}\sum_{j=1}^{n}\|V^Tx_i-V^Tx_j\|^2W_{ij}, \tag{1}$$

where

$$W_{ij}=\begin{cases}\exp\{-\|x_i-x_j\|^2/t\}, & \text{if}\quad y_i=y_j=C_k, k=1,\ldots,c,(t\in\Re)\\ 0, & \text{if}\quad y_i\neq y_j,\end{cases} \tag{2}$$

and the interclass separability is characterized from the penalty graph by the term

$$G_p=\sum_{i=1}^{n}\sum_{j=1}^{n}\|V^Tx_i-V^Tx_j\|^2W'_{ij}, \tag{3}$$

where

$$W'_{ij}=\begin{cases}\exp\{-\|x_i-x_j\|^2/t\}, & \text{if}\quad y_i\neq y_j,(t\in\Re)\\ 0, & \text{if}\quad y_i=y_j.\end{cases} \tag{4}$$

From the definition of $W_{ij}$ and $W'_{ij}$, we know that $\exp\{-\|x_i-x_j\|^2/t\}$ is a strictly monotone decreasing function with respect to the distance between two variables $x_i$ and $x_j$. $W_{ij}$ indicates the importance degree of $x_i$ and $x_j$ (in the same-class) to the intrinsic graph, the smaller the distance $\|x_i-x_j\|^2$ is, the larger the importance degree would be; otherwise, the importance degree is zero. Similarly, $W'_{ij}$ implies the importance degree of $x_i$ and $x_j$ (in the not-same-class) to the penalty graph, the smaller the distance $\|x_i-x_j\|^2$ is, the larger the importance degree would be; otherwise, the importance degree is zero.

To gain more insight into Eq. (1), we rewrite the square of the norm in the form of the matrix trace

$$\begin{aligned}G_c&=\sum_{i=1}^{n}\sum_{j=1}^{n}\|V^Tx_i-V^Tx_j\|^2W_{ij}\\&=\sum_{i=1}^{n}\sum_{j=1}^{n}tr\{(V^Tx_i-V^Tx_j)(V^Tx_i-V^Tx_j)^T\}W_{ij}\\&=\sum_{i=1}^{n}\sum_{j=1}^{n}tr\{V^T(x_i-x_j)(x_i-x_j)^TV\}W_{ij}.\end{aligned} \tag{5}$$

Since the operation of trace is linear and $W_{ij}$ is a scalar, Eq. (5) can be easily simplified as

$$\begin{aligned}G_c&=tr\left\{V^T\left(\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-x_j)W_{ij}(x_i-x_j)^T\right)V\right\}\\&=tr\{V^T(2XDX^T-2XWX^T)V\}\\&=2tr\{V^TX(D-W)X^TV\}\\&=2tr\{V^TXLX^TV\},\end{aligned} \tag{6}$$

where $D=\text{diag}(D_{11},\ldots,D_{nn})$, $D_{ii}=\sum_{j=1}^{n}W_{ij}(i=1,\ldots,n)$ and $L=D-W$ is called the Laplacian matrix. Similar to Eqs. (5) and (6), we can simplify Eq. (3) as follows:

$$G_p=2tr\{V^TX(D'-W')X^TV\}=2tr\{V^TXL'X^TV\}, \tag{7}$$

where $D'=\text{diag}(D'_{11},\ldots,D'_{nn}), D'_{ii}=\sum_{j=1}^{n}W'_{ij}(i=1,\ldots,n)$ and $L'=D'-W'$. In order to enhance the classification result, we try to find the projection axes which will draw the same-class samples closer together while simultaneously making the not-same-class samples distant from each other, which means that the desirable projection axes should minimize the intrinsic graph term $G_c$, meanwhile maximize the penalty graph term $G_p$. So, we can obtain just such projection axes by maximizing the following