Pattern Recognition 60 (2016) 554–570

Contents lists available at ScienceDirect

Pattern Recognition

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

Enhancing information discriminant analysis: Feature extraction with linear statistical model and information-theoretic criteria ☆

Liling Li [a,b], Lan Du [a,b,*], Wei Zhang [a,b], Hua He [a,b], Penghui Wang [a,b]

[a] National Lab of Radar Signal Processing, Xidian University, Xi'an 710071, China
[b] Collaborative Innovation Center of Information Sensing and Understanding at Xidian University, Xi'an 710071, China

ARTICLE INFO

Article history:
Received 10 November 2015
Received in revised form
2 June 2016
Accepted 4 June 2016

Keywords:
Feature transformation
Information theory
Mutual information
Linear statistical model
Classification

ABSTRACT

In this paper, we develop a novel feature transformation method for supervised linear dimensionality reduction. Existing methods, *e.g.*, Information Discriminant Analysis (IDA), estimate the first and second order statistics of the data in the original high-dimensional space, and then design the transformation matrix based on the information-theoretic criteria. Unfortunately, such transformation methods are sensitive to the accuracy of the statistics estimation. To overcome this disadvantage, our method describes the statistical structure of the transformed low-dimensional subspace via a linear statistical model, which can reduce the number of unknown parameters, while simultaneously maximizes the mutual information (MI) between the transformed data and their class labels, which can ensure the between-class separability according to the information theory. The key idea is that we seek the optimal model parameters, including the transformation matrix, via the joint optimization of MI function and log-likelihood function, therefore, this method can not only reduce the estimation errors but also maximize the between-class separability. Experimental results based on synthetic dataset and benchmark datasets demonstrate the better performance of our method over other related methods.

1. Introduction

With the rapid development of science and technology and the emergence of large-scale high-dimensional data, people urgently need to classify high-dimensional data accurately and efficiently. In high-dimensional data, some features are redundant for classification, which not only increase the cost of data processing, but also seriously affect the accuracy of classification. Therefore, dimensionality reduction becomes an important preprocessing step for high-dimensional data classification. There are two major categories of dimensionality reduction methods: feature selection and feature transform [1]. Feature selection methods only keep useful features and discard others. Typical examples include Fisher [2], Relief [3], exhaustive method [2] and so on. Feature transform methods construct new features from the original variables. Typical examples include Principal Component Analysis (PCA) [4], Linear Discriminant Analysis (LDA) [5,6], Median–Mean Line based Discriminant Analysis (MMLDA) [34], Information Discriminant Analysis (IDA) [7], Independent Component Analysis (ICA) [25–30].

In this paper we focus on the feature transformation design for high-dimensional data classification. A feature transform method is usually coupled with an appropriate criterion, which measures the joint "importance" of a set of features. If the criterion is differentiable with respect to the parameters of the transform, and the transform is smooth, it is possible to learn the transform by optimizing the criterion. A well-known feature transform method is LDA, which maximizes the between-class scattering while minimizing the within-class scattering of the transformed data [5]. An advantage of LDA is that the transformation matrix can be found analytically, thereby avoiding numerical optimization. However, the dimensionality of the transformed space in LDA can only be less than the number of data classes, which greatly restricts its performance and applications. To overcome this disadvantage and improve the robustness on outliers, MMLDA is proposed based on the traditional LDA method in [34]. The median mean line is introduced in MMLDA to substitute the original mean vector in LDA, which is the line through the median sample and mean vector. Due to the median sample is nearly stable for the outliers, the median mean line is more robust than the mean vector to outliers. In addition, since the new defined between-class scattering matrix in MMLDA is non-singular for any transformed subspace size, then there is no restriction for the transformed space size any more in the MMLDA method. Besides the

☆This work was partially supported by the National Science Foundation of China (Nos. 61271024 and 61322103), the Foundation for Doctoral Supervisor of PR China (No. 20130203110013), and the Science Foundation of Shaanxi Province (No. 2015JZ016).

* Correspondence to: 2 Taibai Rd., Xi'an, Shaanxi 710071, China.
E-mail address: dulan@mail.xidian.edu.cn (L. Du).

http://dx.doi.org/10.1016/j.patcog.2016.06.004
0031-3203/

improved versions of LDA, some feature transformation methods based on the information-theoretic criteria can also improve the performance of the traditional LDA [8,14]. Tao et al. [8] designs the transformation matrix by maximizing the mutual information (MI) between the transformed data and their class labels [1,7,9–11,24]. According to the theoretic analysis [12,13], the larger the MI is, the tighter the bound of Bayes error rate will be. IDA [7] and the work of [14] are two typically supervised dimensionality reduction methods based on the information-theoretic criteria. Since the MI is very difficult to calculate directly for non-Gaussian distributions, IDA approximates the entropy of the Gaussian mixture model (GMM) with the entropy of a global Gaussian distribution, and provides an approximate expression to the Shannon MI. In the work of [14], an analytic and explicit expression for the gradient of the Shannon MI with respect to the transformation matrix is given for any distribution, thus the Shannon MI can be maximized directly without any approximation. For convenience, the method in [14] is referred to as mutual information maximization (MIM) in this paper. Both methods estimate the first and second order statistics of the data in the high-dimensional space, and then design the transformation matrix based on the information-theoretic criteria. However, in practice, especially in the case of limited data, it is difficult to accurately estimate these statistics in the original high-dimensional space. Thus the transformation matrix obtained may be nonoptimal, which will affect the final classification performance.

This paper proposes a novel supervised dimensionality reduction method based on linear statistical model and information-theoretic criterion, conveniently called enhancing information discriminant analysis (EIDA). Our method describes the statistical structure of the transformed low-dimensional subspace via a linear statistical model, and utilizes the MI based information-theoretic criterion to further maximize the between-class separability. Inspired by IDA, our method seeks the optimal model parameters, including transformation matrix, via the joint optimization of MI function and log-likelihood function. The main contributions of this paper are summarized in the following.

1) In our linear statistical model, the original high-dimensional sample is represented as the linear combination of low-dimensional feature and additive measurement noise. based on such a linear model, the statistics can be estimated in the low-dimensional subspace. therefore, the estimation errors can be reduced due to the lower freedom of unknown parameters, compared with those got from the original high-dimensional data space, especially in the case of limited data. 2) in our optimization algorithm, the joint optimization designs the transformation matrix and estimates the data statistics simultaneously. Therefore, we can not only ensure the between-class separability, but also describe the observed data as accurately as possible.

Here we give Table 1 to list all notations in our EIDA model to help readers to understand the dimensions of matrices and definitions of notations clearly.

The remainder of the paper is organized as follows: Section 2 is a brief overview of Bayesian classification, MI, and IDA. Section 3 introduces EIDA. A complete algorithm is developed and presented in this section with some detailed derivations given in Appendix. Section 4 describes some related methods. In Section 5, we validate the performance of our method on synthetic dataset and some benchmark datasets. Section 6 summarizes the paper by presenting some concluding remarks.

Table 1
List of notations in the EIDA model.

d: Dimensionality of original data
m: Dimensionality of low-dimensional latent feature
N_k: Data size of training sample for class k
n: Number of total samples (include training samples and test samples)
$\boldsymbol{x} \in \mathbb{R}^{d\times 1}$: Original data
$\boldsymbol{y} \in \mathbb{R}^{m\times 1}$: Low-dimensional latent feature of $\boldsymbol{x}$
$\mathbf{A} \in \mathbb{R}^{d\times m}$: Transformation matrix
$\boldsymbol{\epsilon} \in \mathbb{R}^{d\times 1}$: Measurement noise
$\boldsymbol{\Psi}_k \in \mathbb{R}^{d\times d}$: Diagonal matrix of the noise covariance on class k
$\mathbf{F}_k \in \mathbb{R}^{d\times d}$: Prior precision matrix for original data from class k
$\boldsymbol{x}_{ki} \in \mathbb{R}^{d\times 1}$: ith data from class k
$\boldsymbol{y}_{ki} \in \mathbb{R}^{m\times 1}$: Low-dimensional latent feature of $\boldsymbol{x}_{ki}$
$\boldsymbol{\mu}_{xk} \in \mathbb{R}^{d\times 1}$: Prior mean vector for original data from class k
$\boldsymbol{\Sigma}_{xk} \in \mathbb{R}^{d\times d}$: Prior covariance matrix for original data from class k
$\boldsymbol{\mu}_{yk} \in \mathbb{R}^{m\times 1}$: Condition posterior mean vector for latent feature from class k
$\boldsymbol{\Sigma}_{yk} \in \mathbb{R}^{m\times m}$: Condition posterior covariance matrix for latent feature from class k
$\boldsymbol{\mu}_k \in \mathbb{R}^{d\times 1}$, $\boldsymbol{\mu} \in \mathbb{R}^{d\times 1}$: $\boldsymbol{\mu}_k = \mathbf{F}_k\boldsymbol{\mu}_{\boldsymbol{x}k}$, $\boldsymbol{\mu} = \sum_{k=1}^{K} w_k\boldsymbol{\mu}_k$
$\boldsymbol{\Omega}_k \in \mathbb{R}^{d\times d}$, $\boldsymbol{\Omega} \in \mathbb{R}^{d\times d}$: $\boldsymbol{\Omega}_k = \mathbf{F}_k\boldsymbol{\Sigma}_{\boldsymbol{x}k}\mathbf{F}_k$, $\boldsymbol{\Omega} = \sum_{k=1}^{K} w_k\left[\boldsymbol{\Omega}_k + (\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^T\right]$
$\mathbf{O}_k \in \mathbb{R}^{d\times d}$, $\mathbf{O} \in \mathbb{R}^{d\times d}$: $\mathbf{O}_k = \mathbf{A}\left(\mathbf{A}^T\boldsymbol{\Omega}_k\mathbf{A}\right)^{-1}\mathbf{A}^T$, $\mathbf{O} = \mathbf{A}\left(\mathbf{A}^T\boldsymbol{\Omega}\mathbf{A}\right)^{-1}\mathbf{A}^T$

2. Background

2.1. Bayesian classification

The Bayesian classifier [2] is often used to evaluate the performance of a supervised dimensionality reduction method. For a discrete-valued class variable C and a continuous feature variable $\boldsymbol{Y} \in \mathbb{R}^{m\times 1}$, let $p(c = k) \triangleq w_k$ be the prior distribution for $\forall\, k = 1, 2, \ldots, K$ with K denoting the number of classes, and $p(\boldsymbol{y}|k) = N\left(\boldsymbol{y}; \boldsymbol{\mu}_{yk}, \boldsymbol{\Sigma}_{yk}\right)$ be the class-conditional probability density function (PDF), where $\mathrm{N}\left(\boldsymbol{y}; \boldsymbol{\mu}_{yk}, \boldsymbol{\Sigma}_{yk}\right)$ denotes the Gaussian distribution for variable $\boldsymbol{y}$ with mean vector $\boldsymbol{\mu}_{yk}$ and covariance matrix $\boldsymbol{\Sigma}_{yk}$. According to Bayes rule, the posterior probability $p(k|\boldsymbol{y})$ can be expressed as $p(k|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|k)w_k}{p(\boldsymbol{y})}$, where $p(\boldsymbol{y}) = \sum_{k=1}^{K} p(\boldsymbol{y}|k)w_k$. The Bayesian classifier will assign $\boldsymbol{y}$ to class t if

$$t = \arg\max_{k:1,2,\ldots,K} p(k|\boldsymbol{y}) \tag{1}$$

The above formula can be further expressed as

$$\begin{aligned} t &= \arg\max_{k:1,2,\ldots,K} \ln\left[p(\boldsymbol{y}|k)w_k\right] \\ &= \arg\max_{k:1,2,\ldots,K}\left[-\frac{m}{2}\ln(2\pi) - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_{yk}\right| - \frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{\mu}_{yk}\right)^T\boldsymbol{\Sigma}_{yk}^{-1}\left(\boldsymbol{y} - \boldsymbol{\mu}_{yk}\right) + \ln(w_k)\right] \end{aligned} \tag{2}$$

where T denotes the transpose of a vector or a matrix. If the covariance matrices for all of the classes are identical, *i.e.*, $\boldsymbol{\Sigma}_{yk} = \boldsymbol{\Sigma}_y$, $\forall\, k = 1, 2, \ldots, K$. In this case, from Eq. (2), we have

$$t = \arg\max_{k:1,2,\ldots,K}\left[\boldsymbol{\mu}_{yk}^T\boldsymbol{\Sigma}_y^{-1}\boldsymbol{y} - \frac{1}{2}\boldsymbol{\mu}_{yk}^T\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\mu}_{yk} + \ln(w_k)\right] \tag{3}$$

where the resulting discriminant functions are linear. A Bayesian classifier that uses linear discriminant functions is called a linear Bayesian classifier.

Another simple case arises when the covariance matrices are different for each category, then, Eq. (2) can be written as

$$t = \arg\max_{k:1,2,\ldots,K}\left\{-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{\Sigma}_{yk}^{-1}\boldsymbol{y} + \boldsymbol{\mu}_{yk}^T\boldsymbol{\Sigma}_{yk}^{-1}\boldsymbol{y} - \frac{1}{2}\boldsymbol{\mu}_{yk}^T\boldsymbol{\Sigma}_{yk}^{-1}\boldsymbol{\mu}_{yk} - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_{yk}\right| + \ln(w_k)\right\} \tag{4}$$