Neurocomputing 207 (2016) 346–353

Contents lists available at ScienceDirect

Neurocomputing

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

Approximative Bayes optimality linear discriminant analysis for Chinese handwriting character recognition

Chao Yao, Gong Cheng *

Department of Control and Information, School of Automation, Northwestern Polytechnical University, 127 Youyi Xilu, Xi'an 710072, PR China

ARTICLE INFO

Article history:
Received 3 December 2015
Received in revised form
17 February 2016
Accepted 4 May 2016
Communicated by Feiping Nie
Available online 13 May 2016

Keywords:
Dimension reduction
Bayes optimality
Chinese handwriting character recognition
Large category
Class separation problem

ABSTRACT

Discriminant subspace learning is an important branch for pattern recognition and machine learning. Among the various methods, Bayes optimality linear discriminant analysis (BLDA) has shown its superiority both in theory and application. However, due to the computational complexity, BLDA has not been applied to large category pattern tasks yet. In this paper, we propose an approximative Bayes optimality linear discriminant analysis (aBLDA) method for Chinese handwriting character recognition, which is a typical large category task. In the aBLDA, we first select a set of convex polyhedrons that are obtained by the state-of-the-art methods, then the searching zones are limited to these polyhedrons. Finally, the best of them is chosen as the final projection. In this way, the computational complexity of BLDA is reduced greatly with comparable accuracy. To find more than 1D projections, the orthogonal constraint is employed in the proposed method. The experimental results on synthetic data and CASIA-HWDB1.1 show the effectiveness of the proposed method.

1. Introduction

Extracting efficient feature from high-dimensional data is a crucial procedure in pattern classification tasks, for it could not only reduce the computational complexity, but also lead to better performance by removing the redundancy and noise of the data, meanwhile overcoming the curse-of-dimension problem for statistical classifier. In the past years, lots of methods have been proposed to solve the problem, such as Principal Component Analysis (PCA) [1], Linear Discriminant Analysis (LDA) [2], Independent Component Analysis (ICA) [3], Locality Preserving Projections (LPP) [4], kernel PCA [5], kernel LDA [6], ISOMAP [7], Locally Linear Embedding (LLE) [8] and so on. These methods could be categorized into linear and nonlinear ones based on different viewpoints, or supervised and unsupervised ones based on the samples' label information.

For large category pattern classification tasks, linear supervised feature extraction methods are widely employed. LDA, which aims to minimize the within class scatter and maximize the between class scatter at the same time in the low-dimensional space, is the most popular one among those methods. LDA only calculates the between scatter matrix and the within scatter matrix, then solves the eigenvalue decomposition problem, which makes LDA be implemented easily and computed efficiently. However, LDA would overemphasize on the classes that locate far-apart with each other for multi-class task. Various methods were developed to overcome the problem. Lotlikar et al. [9] introduced a fractional step method for the problem. In their method, a weighting function was used to increase the weights of the pairs that would be merged in fractional steps. Loog et al. [10] proposed another weighting function using the approximation of pairwise Bayes function. The function gave the pairs that locate nearly a large value according to Bayes rules. Recently, Zhang et al. [11] proposed a new weighting function based on the confusion matrix for the specific classifier. Nevertheless, these methods are not directly related to the classification error, and still have room for improvement in theoretical perspective.

Designing a Bayes optimal criterion for multi-class dimension reduction is a well known difficult job. However, progresses have been achieved under some constraint. For multi-class homoscedastic Gaussian distributed data with equal prior, Geisser [12] gave the formulation of Bayes error. Schervish [13] solved the 2D data for three classes which is a special case of [12]. Hamsici et al. [14] proposed a Bayes optimality linear discriminant analysis (BLDA). They first demonstrated the projections, which projected the means to the same order, formed a convex set. Therefore, the convex optimization methods could be used to solve the problem. To obtain the 2D or higher dimensional subspace, they iteratively found the 1D Bayes optimal projection in the orthogonal space. BLDA has shown its superiority to the other methods in various data sets. However, Hamsici claimed that $C!/2$ convex problems should be solved for

* Corresponding author.
E-mail addresses: yaochao@nwpu.edu.cn (C. Yao), chenggong1119@gmail.com (G. Cheng).

http://dx.doi.org/10.1016/j.neucom.2016.05.017
0925-2312/

each projection at maximum, where C was the number of classes. With C getting large, the number of convex spaces $C!/2$ increases sharply. Although they linearly approximated the Bayes error function to reduce the computational complexity, the performance of this approximation was not promising.

We focus on the handwriting Chinese character recognition (HCCR) task in this paper, which is a typical large category problem. For HCCR, LDA is a popular dimension reduction method. However, it might merge the similar characters (characters that look alike) in the subspace, which degenerates the performance of HCCR. Methods on two aspects are proposed to solve this problem. On one hand, extra features are extracted for the similar characters. Leung et al. [15] extracted extra features for similar characters in the critical regions. Gao et al. [16] adopted LDA to obtain more discriminant feature for similar character pairs, then used a compound distance metric to classify the similar characters. Considering that the similar characters were not only pair-wised, Wang et al. [17] extracted extra features for several similar characters simultaneously. In these methods, an extra classifier is employed to classify the similar characters, which leads to extra computational time and storage space. On the other hand, similar characters must be the classes located nearly in the feature space. The aforementioned weighting methods could improve the separation problem of LDA. Zhang et al. [18] proved that the family of weighting LDA methods could improve the performance of HCCR, this is mainly because these methods assign large weights to the similar character pairs. However, to our best knowledge, the Bayes optimal methods have not been applied to HCCR or other large category pattern classification tasks yet due to their heavy computational complexity.

In this paper, we present an approximative BLDA (aBLDA) for HCCR. The main contribution of our work is that we reduce the searching spaces of BLDA significantly so that it could be used in HCCR task efficiently with comparable accuracy. In the proposed method, we do not search all the possible convex sets but those found by the popular linear dimension reduction methods, such as LDA, aPAC, the POW family and other methods that are already adopted by large category tasks. We first apply a set of dimension reduction methods on HCCR to get their individual 1D subspaces, and then use the 1D subspaces to find the convex sets determined by the above methods. Finally, the local Bayes optimality projections are found on these convex sets and the best of them is chosen as the final result. More projections could be obtained iteratively in the orthogonal space by our method. The proposed approximative BLDA is expected to have the following advantages: (i) the number of convex sets that need to be searched is far less than BLDA; (ii) with the improvement of Bayes optimality, the performance would be better than that of the existed ones. The experiments on both the synthetic data and CASIA-HWDB1.1 demonstrate the validity of the proposed method.

The rest of the paper is organized as follows: the idea of BLDA is introduced in Section 2, then details of the approximative BLDA are presented in Section 3. We perform experiments in Section 4 to show the effectiveness of the proposed method and finally conclude it in Section 5.

2. Review of Bayes optimality in linear discriminant analysis

In this section, we will first introduce the details of BLDA [14]. For convenience, we present the important notations in the paper in Table 1. In the paper, uppercase and lowercase letters in boldface denote matrix and vector. Matrix dimensions are shown as $(m \times n)$, where m and n are the numbers of rows and columns, respectively.

Table 1
Notations.

Notation	Description
$\mathbf{X}$	Data matrix
μ_i	Centroid of the ith class
μ	Centroid of all samples
$\mathbf{S}_b$	Between-class scatter matrix
$\mathbf{\Sigma}_i$	Covariance matrix of the ith class
$\mathbf{S}_w$	Within-class scatter matrix
$\mathbf{S}_t$	Total scatter matrix
n	Number of training samples
n_i	Number of training samples in the ith class
C	Number of classes
d	Number of dimensions
l	Number of retained dimensions obtained by feature extraction

2.1. The Bayes optimal 1D subspace

Let $\mathbf{X} \in \mathbb{R}^{d \times n}$, where each row of $\mathbf{X}$ is a sample, be drawn from a set of homoscedastic Gaussian distributions with different means μ_i and equal prior probabilities, where $\mathbf{\Sigma}_i = \mathbf{\Sigma}, \forall i = 1, 2, \ldots, C$. Then the Bayes error when projecting the data on a single dimension can be formulated as:

$$B_{error} = 2C^{-1} \sum \Phi\left(\frac{\eta_{(1)} - \eta_{(i+1)}}{2}\right) \quad (1)$$

where $\Phi(\cdot)$ is the cumulative distribution function (cdf) of a standard Normal distribution, η_i are the ordered projected and whitened means, where $\eta_{(1)} < \eta_{(2)} < \cdots < \eta_{(C)}$, and

$$\eta_i = \frac{\mathbf{v}^T \mu_i}{\sqrt{\mathbf{v}^T \mathbf{\Sigma} \mathbf{v}}} = \mathbf{v}^T \hat{\mu}_i \quad (2)$$

To solve the above problem, Hamsici proves that a constraint region $\mathcal{A}$, where all vectors $\mathbf{v}$ sampled from it generate the same ordered $\{\eta_{(i)}\}$, is a convex polyhedron, and the Bayes error function $B(\mathbf{v})$ is convex for $\forall \mathbf{v} \in \mathcal{A}$. Details of the proof could be found in [14]. In this way, the problem could be solved by searching all the convex regions and choosing $\mathbf{v}_{opt}$ according to the minimal $B(\mathbf{v})$.

2.2. Linear approximative BLDA

In [14], a linear approximative BLDA (lBLDA) method is introduced. In the method, Eq. (1) is replaced by $\frac{1}{C}\sum a(\eta_{(i)} - \eta_{(i+1)}) + b$, then Eq. (1) becomes:

$$B_{error} = -\frac{a(\eta_{(i)} - \eta_{(C)}) + (C-1)b}{C} \quad (3)$$

For $\eta_i = \mathbf{v}^T \hat{\mu}_i$, where $\hat{\mu}_i$ is the whitening mean of the ith class. The solution for Eq. (3) is $\mathbf{v} = \frac{\hat{\mu}_{ij}}{\|\hat{\mu}_{ij}\|}$, where $\hat{\mu}_{ij}$ is the maximal of the $C(C-1)/2$ distances between the whitening means of each pair of classes. In this way, we just need to compute the $C(C-1)/2$ distances and the computational complexity is far reduced.

3. Approximative Bayes optimality linear discriminant analysis

We will discuss the obstacles for applying BLDA to the large category tasks here. To find 1D optimal Bayes projection, the convex problem defined in Eq. (1) should be solved for each constraint polyhedron. The number of the polyhedrons is related to the practical situation and its upper boundary is $C!/2$. To find l-dimensional data, the above procedure should be repeated l times. For the large category problem, the class number C is always very large, which makes the number of search spaces be extremely large. Thus, BLDA could not be used in such pattern recognition tasks directly.