Neural Networks 34 (2012) 56–64

Contents lists available at SciVerse ScienceDirect

Neural Networks

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

Maxi-Min discriminant analysis via online learning

Bo Xu [a], Kaizhu Huang [b,*], Cheng-Lin Liu [b]

[a] Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun East Road Beijing 100190, PR China

[b] National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun East Road Beijing 100190, PR China

ARTICLE INFO

Article history:
Received 17 June 2011
Received in revised form 8 May 2012
Accepted 12 June 2012

Keywords:
Linear discriminant analysis
Dimensionality reduction
Multi-category classification
Handwritten Chinese character recognition

ABSTRACT

Linear Discriminant Analysis (LDA) is an important dimensionality reduction algorithm, but its performance is usually limited on multi-class data. Such limitation is incurred by the fact that LDA actually maximizes the average divergence among classes, whereby similar classes with smaller divergence tend to be merged in the subspace. To address this problem, we propose a novel dimensionality reduction method called Maxi-Min Discriminant Analysis (MMDA). In contrast to the traditional LDA, MMDA attempts to find a low-dimensional subspace by maximizing the minimal (worst-case) divergence among classes. This "minimal" setting overcomes the problem of LDA that tends to merge similar classes with smaller divergence when used for multi-class data. We formulate MMDA as a convex problem and further as a large-margin learning problem. One key contribution is that we design an efficient online learning algorithm to solve the involved problem, making the proposed method applicable to large scale data. Experimental results on various datasets demonstrate the efficiency and the efficacy of our proposed method against five other competitive approaches, and the scalability to the data with thousands of classes.

1. Introduction

Dimensionality reduction has been an important topic in machine learning and pattern recognition. Principal Component Analysis (PCA) (Gao, 2008) does not guarantee the discrimination performance as it does not consider the label information. Linear Discriminant Analysis (LDA), developed by Fisher in 1936, is a popular method that has achieved great success in many fields (Fukunaga, 1990). Under the homoscedastic Gaussian assumption, LDA is equivalent to finding the maximum-likelihood (ML) parameter estimates and leads to the optimal projection axis used for two-category data (Campbell, 2008). When applied to multi-category (e.g., c-category) data, LDA can still achieve good performance in many cases. Precisely speaking, Rao (1948) showed that $c-1$ dimensional subspace given by LDA, wherein c is the class number, and is also guaranteed to be Bayes optimal in multi-class homoscedastic case under the condition that the data features $d \geq c$. Fig. 1(a) is one example where LDA can find the good projection axis for three-class separation.

However, LDA may fail to find good projection for other multi-class data, especially when the category number is far larger than the data features (Loog, Duin, & Haeb-Umbach, 2001), e.g., in Chinese character recognition (with 3755 classes and merely several hundred features). In this case, it is impossible to reduce the dimensionality to any number equal to or slightly smaller than $c-1$. Fig. 1(b) illustrates a typical example that LDA fails to find a good projection when the dimensionality is reduced to 1 for a three-class problem. Clearly, by LDA, the transformed data of class 1 and class 2 would overlap with each other heavily, leading to worse performance for consequent classification. This problem of LDA, or more clearly, the phenomenon that LDA tends to merge similar or closer classes when the dimension of the projected subspace is strictly lower than the class number minus one, is called the class separation problem in the literature (Tao, Li, Wu, & Maybank, 2009). In contrast, the dashed axis in Fig. 1(b), would be a reasonable projection axis that can appropriately make the data of each class well separated.

The criterion of LDA is trying to search a low-dimensional subspace which can maximize the between-class covariance while minimizing the within-class covariance. Using Lemma 1 (provided in Section 3.2), LDA actually exploits an *average* setting, i.e., LDA tries to maximize the *average* divergences among different classes. The divergence of any two classes is defined as the distance between the mean vectors of the two classes in the whitening space. To maximize the *average* divergence, LDA tends to find the subspace preserving the larger divergences and ignoring the smaller divergences, as illustrated in Fig. 1(b). This causes the overlap of the similar classes, with smaller divergences, after data transformation.

To address the class separation problem of LDA, there have been several proposals in the literature. Loog et al. (2001)

* Corresponding author.
E-mail addresses: kzhuang@nlpr.ia.ac.cn, kaser.huang@gmail.com (K. Huang).

0893-6080/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2012.06.001

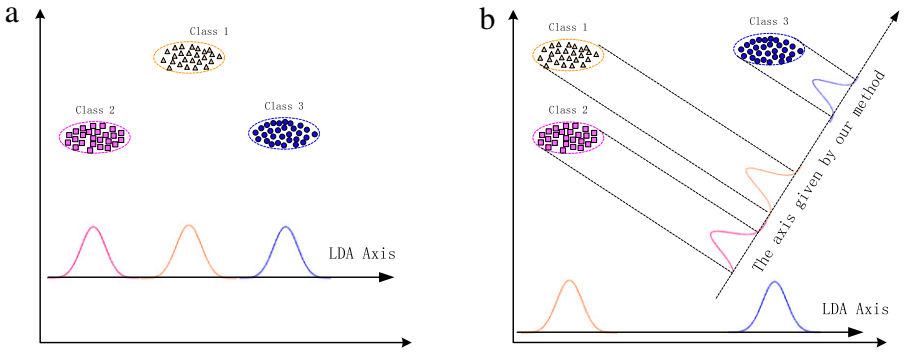

Fig. 1. Illustration of LDA for multi-class data.

developed a heuristic method called approximate Pairwise Accuracy Criterion (aPAC) that adds larger weights for similar classes in the estimation of the between-class covariance. Lotlikar and Kothari (2000) proposed the so-called Fractional-step Linear Discriminant Analysis (F-LDA) by heuristically and iteratively reducing dimension from a high-dimensional space to the low-dimensional space. Recently, Abou-Moustafa, de la Torre, and Ferrie (2010) designed the Pareto Discriminant Analysis (PDA) by forcing the pairwise distance to be equal after transformation. These methods usually deal with the class separation problem by imposing different weights on classes, either iteratively or directly. However, the weighting function is always ad-hoc and often needs to be adapted in different applications. For PDA, the involved optimization problem is non-convex, making its performance usually limited in practice. More related work can be referred to Section 2.

Unlike previous approaches, in this paper, a novel *worst-case* framework called Maxi-Min Discriminant Analysis (MMDA) is proposed. More specifically, instead of maximizing the *average* divergence among different classes, MMDA attempts to maximize the *minimal* (worst-case) divergence. In this worst-case setting, MMDA tries to push away each pair of classes with small divergence as large as possible. This consequently avoids the aforementioned problem and hence presents a more rigorous method (compared with aPAC and F-LDA). Obviously, the proposed MMDA method is still optimal for two-class problems under the homoscedastic Gaussian assumption, since it is degraded to the standard LDA when the class number is equal to two. Hence, the proposed worst-case method can be seen as a more generalized version of LDA for multi-class problems.

One important contribution of this paper is that we formulate the MMDA problem as a convex programming problem, or more precisely a Semi-Definite Programming (SDP) problem. Since SDP is computationally intractable even for medium-size data, we first transform the involved SDP problem to a large margin problem and then present an efficient online learning algorithm to solve it. The proposed online algorithm is important in that (a) it is computationally more efficient by removing the constraint of SDP, and (b) it has a nice convergence property. We note that Bian and Tao (2010) and Yu, Jiang, and Zhang (2011) proposed a similar model from the view point of distance metric learning or dimensionality reduction. However, their models are basically SDP problems and are hence intractable for large-scale data.

The paper is organized as follows. In Section 2, we detail the related work. In Section 3, we present our novel worst-case framework for dimensionality reduction in detail. The model definition, the theoretical justification and practical optimization will be discussed in turn in this section. In Section 4, we evaluate our algorithm and report experimental results. Finally, we set out the conclusion with some remarks.

A preliminary version of this paper has been early published in Xu, Huang, and Liu (2010), which is however significantly expanded both in theory and experiments in the current version.

2. Related work

There are a number of dimensionality reduction approaches related to our work (Abou-Moustafa et al., 2010; He & Niyogi, 2003; Loog et al., 2001; Lotlikar & Kothari, 2000; Sugiyama, 2006; Tang, Suganthan, Yao, & Qin, 2005; Tao et al., 2009; Yan et al., 2007).

Sugiyama (2006, 2007) developed the Local Fisher Discriminant Analysis (LFDA) method that combines the merits of Locality Preserving Projection (LPP) (He & Niyogi, 2003) into LDA. This method is shown to be very promising in many real datasets. However, it is mainly designed for solving classification tasks when classes distributions are multi-model and its performance is limited in handling the class separation problem. Loog et al. (2001) developed a method called approximate Pairwise Accuracy Criterion (aPAC) that adds larger weights for similar classes in the estimation of the between-class covariance. This method is well motivated and partially solves the class separation problem. However, it remains a problem how to select an optimal weighting function. Tang et al. (2005) proposed a relevance weighted LDA which incorporates the inter-class relationships as relevance weights into the estimation of the overall within-class scatter matrix in order to improve the performance of the basic LDA method. The major problem is still how to select the optimal weighting function. Another related approach called Fractional-step Linear Discriminant Analysis (F-LDA) is proposed in Lotlikar and Kothari (2000). F-LDA is a heuristic method, which can generate better classification accuracy by iteratively reducing dimension from a high-dimensional space to the low-dimensional space. This improves the robustness of choosing the weighting function. Its performance is limited in that a large number of steps should be used to collapse each dimension and the choice of a scaling parameter is always critical for the final result. Marginal Fisher Analysis (MFA) (Yan et al., 2007) is also highly related to our method. MFA characterizes the interclass compactness by the neighboring points in the same class and characterizes the interclass separability by the connection of marginal points. However, graph construction based on the whole data-set is time-consuming, which limits its application.

As a short summary, the above methods usually deal with the class separation problem by imposing different weights on classes, either iteratively or directly. However, the weighting function is always ad-hoc and often needs to be adapted in different applications. Recently, Pareto Discriminant Analysis was proposed to address the class-separation problem (Abou-Moustafa et al., 2010) by forcing the pairwise distance to be equal after transformation. There are two shortcomings for the method. On the one hand, it involves a non-convex programming problem; on the other hand, the so-called Pareto optimal criterion may essentially not be a good criterion because a bad local minimum can be a Pareto optimal point as well. Yu et al. (2011) proposed a similar criterion called minimal distance maximization for