Neurocomputing 173 (2016) 154–162

Contents lists available at ScienceDirect

Neurocomputing

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

An improved Fisher discriminant vector employing updated between-scatter matrix

Chao Yao [a], Zhaoyang Lu [a], Jing Li [a], Wei Jiang [a], Jungong Han [b,*]

[a] State Key Laboratory of Integrated Service Networks, Xidian University, Xian 710071, China
[b] Department of Computer Science and Digital Technologies, Northumbria University, Newcastle upon Tyne NE1 8ST, UK

ARTICLE INFO

Article history:
Received 9 August 2014
Received in revised form
30 October 2014
Accepted 16 November 2014
Available online 2 September 2015

Keywords:
LDA
Classification
Dimension reduction
Updated between-scatter matrix

ABSTRACT

Discriminant analysis is an important and well-studied algorithm in pattern recognition area, and many linear discriminant analysis methods have been proposed over the last few decades. However, in the previous works, the between-scatter matrix is not updated when seeking the discriminant vectors, which causes redundancy for the well separated pairs. In this paper, a between-scatter matrix updating scheme is proposed based on the separable status of the obtained vectors. In our scheme, separable status determination of obtained vectors is decisive. Here, we notice that appropriate separation of a multi-dimensional feature (with homoscedastic Gaussian distribution) may help to find better discriminant vectors, and the separability of a multi-dimensional feature can be deduced from the separability of its elements. To make the discriminant vectors statistically uncorrelated, the algorithm is applied to the S_t-orthogonal space of the obtained vectors in an iterative way. We also extend our method to more general cases, like heteroscedastic distributions, by an appropriate kernel function. Experimental results on multiple databases demonstrate the effectiveness of the proposed method.

1. Introduction

Big data is a hot term describing the availability and the rapid growth of data in recent years. Big data combined with high-powered analytics can lead to more accurate decision making. However, using big data techniques in a real-world application is not that straightforward. To this end, three challenging problems have to be addressed, which are data acquisition, data processing, and data storage. One of the solutions is to reduce the data volume using representative data samples. Therefore, dimension reduction becomes an essential step in many big data applications.

Many solutions have been proposed for reducing the data dimension, such as Principal Component Analysis (PCA) [1], Linear Discriminant Analysis (LDA) [2], Local Discriminant Embedding (LDE) [3]. Among these methods, LDA is one of the most prominent methods due to its simplicity and better performance. LDA is first proposed by Fisher for a two-class problem, and is extended to solve a multi-class problem by Rao [4] later on. The basic idea of LDA is to find the optimal projecting directions that minimize the within class scatter and maximize the between class scatter simultaneously. LDA has been successfully deployed in many applications including image recognition [5,6], data analysis [7,8], *visual recognition* [9,10] and so on. However, there are *at least* three remaining problems in LDA and its variations.

The first one is a non-optimal problem for the multi-class case if the final dimensionality l is strictly smaller than the class number C minus one. This is due to the fact that LDA tends to merge the classes that locate closely. To solve this problem, many algorithms have been proposed in the last decade. Loog et al. [11] propose a weighting function which is the approximation of a pairwise Bayes function. It assigns a relatively large value to the pairs that locate nearly based on Bayes rules. Bian and Tao [12] develop a method that aims to maximize the minimum pairwise distance. The distance between the classes that may potentially be merged is maximized. Recently, we propose a subset method [13], which divides the whole set into several subsets such that the linear discriminant methods can be applied on each subset individually. Our method significantly improves the current linear feature extraction methods, especially in the low-dimensional representations.

The second one is the singularity problem for the within-class scatter matrix in case only a Small Sample Size (SSS) training set is available. The singularity makes the inversion computation of within-class scatter matrix impossible. A simple solution to deal with this problem is called regularized method [14,15]. It adds a small positive number to the diagonal elements of the within-class scatter matrix so that the inversion of the matrix can be guaranteed. Methods based on PCA family are also developed to solve the

* Corresponding author.
E-mail addresses: yaochao@mail.xidian.edu.cn (C. Yao), zhylu@xidian.edu.cn (Z. Lu), jinglixd@mail.xidian.edu.cn (J. Li), jwmianzu@gmail.com (W. Jiang), jungonghan77@gmail.com (J. Han).

http://dx.doi.org/10.1016/j.neucom.2014.11.102
0925-2312/

problem [16]. In these methods, PCA is first performed to remove the null space of the within-class scatter matrix, and then LDA is used in the lower dimensional PCA subspace. Moreover, another branch of methods is based on a pseudo-inversion [17] scheme, in which the pseudo-inversion of within scatter matrix makes LDA applicable.

The third one is to find a suitable constraint that formulizes the relation between discrimination vectors. Generally speaking, there are two typical LDA algorithms: one is Foley–Sammon linear discriminant analysis (FSLDA) [18,19], and the other one is the uncorrelated linear discriminant analysis (ULDA) [20,21]. They both aim to find vectors $\varphi_1, \varphi_2, \ldots, \varphi_n$ that maximize the Fisher criterion iteratively based on different constraints. To do so, different matrix decomposition methods are proposed. Furthermore, it is proven that ULDA always outperforms FSLDA, thanks to its uncorrelated property [22]. Jin et al. [23] also show that ULDA is equivalent to LDA if the eigenvalues of $S_w^{-1}S_b$ are not the same. Here, the uncorrelated constraint plays an important role, and it is successfully applied to other discriminant methods [24,25].

A lot of efforts have been devoted to these three areas and great progresses have been made during last several decades. However, to our best knowledge, the existing works do not take the separability of the feature vectors into consideration during the recursive discriminant analysis procedure. In this paper, our idea is to focus the discriminant analysis on the classes that are not well separated. It is not even necessary to apply the discriminant analysis for the classes that can already be separated from the previous loops and will not be merged in higher dimensional feature space. This is different from LDA that maximizes the distance of all classes for each projecting vector, which causes redundancy. Implementing our idea requires a scheme that is able to accurately and quickly determine the separable status of feature vectors. Here, we build our scheme upon an observation that the multi-dimensional feature with homoscedastic Gaussian distribution must be separable if any of its elements (or variables in feature selection methods) is separable. On the basis of this, an improvement for LDA is proposed then, in which the between-scatter matrix is updated during the procedure. To make the discrimination vectors statistically uncorrelated, we iteratively apply our algorithm to the S_t-orthogonal space of the obtained vectors. The method is also extended to more general cases, like heteroscedastic distributions by kernel mapping function.

The rest of the paper is arranged as following: Section 2 discusses the relation of discriminant capability between the element feature and the multi-dimensional feature vector. We then review LDA and the constraint of the projecting vectors in Section3. Afterwards the improving LDA method is presented in Section 4. Experiments are shown in Section 5, and finally we conclude the proposed method in Section 6.

2. Relation of discriminant capability between multi-dimensional feature and its elements

A simple example is presented in Fig. 1, where the elements are separable in the 1D space of axis X and completely overlap in the 1D space of axis Y. Moreover, for the feature vector in the 2D plane, it is well separated. In this section, we provide the theoretical analysis for the relation between multi-dimensional feature discriminant capability and the discriminant capability of its elements. The intention is to investigate whether the characteristics of the low dimensional elements can be employed to approximate the characteristics of the entire high dimensional feature vector or not.

For two classes of homoscedastic Gaussian data $G_1(\mu_1, \Sigma)$ and $G_2(\mu_2, \Sigma)$, the decision boundary under Bayes' rule is $(\mu_1 - \mu_2)^t \Sigma^{-1}(X_b^t - (\mu_1+\mu_2)/2)$. Thus, $\forall x_b \in X_b, p_{G_1}(x_b) = p_{G_2}(x_b)$. Bayes'

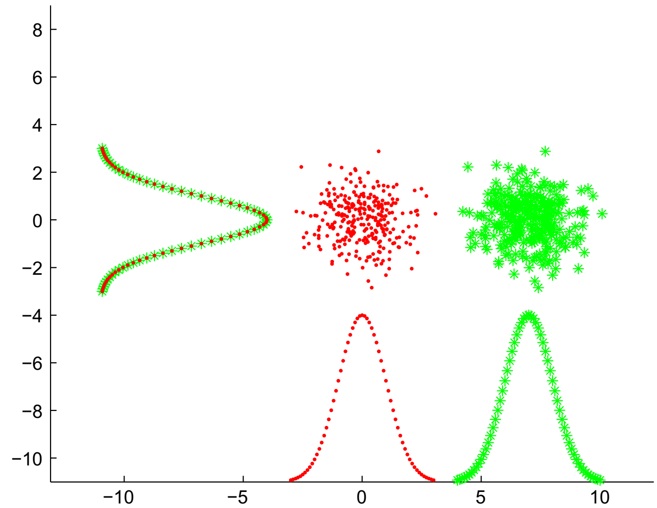

Fig. 1. An example illustrates the discriminant ability between feature and its elements.

error under the assumption is $P(error) = \int \min(G_1P_1, G_2P_2)dx$, where P_1 and P_2 are the priors of the two classes, respectively. For $P(error) \le (P_1V_1 + P_2V_2)\sup\ p(x_b)$, V_1 and V_2 are the volumes of class1 and class2 determined by $\min(G_1P_1, G_2P_2)$, respectively. If the $\sup\ p(x_b) \le \varepsilon$, then Bayes' error $P(error) \le \tau$. Thus, a small ε must lead to a very small τ. For this case, the two classes are deemed to be separable. In a word, we consider the two classes that are separable if the upper probability boundary in the decision plane is very small. With the above definition, we have the following theorem:

Theorem 1. *For two whiten Gaussian data with equal prior, the feature must be separable if any of its element is separable.*

Proof. For the Gaussian data $G_1(\mu_1, \Sigma)$, $\sup\ p_{G_1}(x_b) = \sup(1/(2\pi)^{d/2}|\Sigma|^{-1})e^{-(x_b-\mu_1)^t\Sigma^{-1}(x_b-\mu_1)/2} \le \varepsilon \Longleftrightarrow \inf\ d_m^{G_1}(x_b) = \inf\ (x_b - \mu_1)^t\Sigma^{-1}(x_b-\mu_1)/2 \ge d$. To find the minimum point of $d_m^{G_1}(x_b)$, the following equation holds

$$\inf\ d_m^{G_1}(x_b) = \begin{cases} \text{minimize } (x-\mu_1)^t\Sigma^{-1}(x-\mu_1) \\ \text{subject to } (\mu_1-\mu_2)^t\Sigma^{-1}(x - \frac{\mu_1+\mu_2}{2}) = 0 \end{cases} \quad (1)$$

The gradient of Eq. (1) is

$$2\Sigma^{-1}(x-\mu_1) - 2\frac{(x-\mu_1)^t\Sigma^{-1}\Sigma^{-1}(\mu_1-\mu_2)^t}{\|\Sigma^{-1}(\mu_1-\mu_2)^t\|_2^2} \cdot \Sigma^{-1}(\mu_1-\mu_2)$$

With $x = (\mu_1+\mu_2)/2$, the above equation reaches zero, which means $x = (\mu_1+\mu_2)/2$ is the minimum point of Eq. (1). Thus with the whiten assumption, we have

$$d_m^{G_1}\left(\frac{\mu_1+\mu_2}{2}\right) = \sum_i \frac{|\frac{\mu_1^i - \mu_2^i}{2}|^2}{2}$$

where μ_1^i and μ_2^i are the ith element of μ_1 and μ_2 respectively. If an element feature $(|\mu_1^i - \mu_2^i/2|^2)/2 \ge d$, then $d_m^{G_1}((\mu_1+\mu_2)/2) \ge d$. This helps us to draw a conclusion: for the binary whiten Gaussian data with equal prior, if any of its elements is separable, the feature must be separable.□

Our main result given above provides an insight into the relation between the feature and its element in a classification. The obtained relation is based on homoscedastic Gaussian distribution with equal prior for two-class situation, and Hamsici and Martinez [26] prove that LDA achieves Bayes optimal under such constraints. This conclusion could be used to improve the performance of LDA, which maximizes the distances of all the pair of