Engineering Applications of Artificial Intelligence 43 (2015) 15–26

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

On the linear discriminant analysis for large number of classes

Yizhen Huang [a], Yepeng Guan [a,b,*]

[a] School of Communication and Information Engineering, Shanghai University, China
[b] Key Laboratory of Advanced Displays and System Application, Ministry of Education, China

ARTICLE INFO

Article history:
Received 9 July 2013
Received in revised form
24 April 2014
Accepted 11 March 2015
Available online 22 April 2015

Keywords:
Linear Discriminant Analysis
Trace-ratio
Subspace learning
Face recognition
Pattern recognition

ABSTRACT

We study the challenging problem to classify samples into a large number of classes, and propose the idea of using different Dimensionality-Reduction (DR) projections for different classes of samples. Based on this intuitive idea, the traditional Linear Discriminant Analysis (LDA) and the trace-ratio LDA are formulated to their corresponding new multi-subspace objectives. We justify that certain effects of class-adaptive feature selection are naturally achieved via our multi-subspace DR methods. Experiments on seven datasets show that, our multi-subspace trace-ratio LDA outperform its ratio-trace and single-subspace counterparts, and its advantage is more apparent when the number of classes to be classified is large.

1. Introduction

Linear Discriminant Analysis (LDA) (Belhumeur et al., 1997; Duda et al., 2000; Fukunaga, 1991) is a family of very widely used supervised Dimension Reduction (DR) techniques in the statistics research areas. The original Fisher's LDA (Fisher, 1936) only finds one projection direction u (*i.e.* a vector) for discriminating two classes, and its objective, namely the Fisher's criterion, is to maximize the ratio of between-class scatter matrix S_b to within-class scatter matrix S_w:

$$J(u) = \max_u \frac{|u^T S_b u|}{|u^T S_w u|} \tag{1}$$

Note that, here u is a vector so $u^T S_b u$ and $u^T S_w u$ are two scalars. Thus, it is legitimate to use the term "ratio" here, and $|\cdot|$ is the absolute value operator. In statistical applications, *e.g.* face recognition and image annotation, we usually have a large number of image classes, hence the multi-class LDA is more desired. When r ($r>1$) projection directions U (*i.e.* a matrix) are needed, both $U^T S_b U$ and $U^T S_w U$ are r by r matrices, and their ratio cannot be computed directly. In the traditional LDA, typically the determinant ratio is used:

$$J(U) = \max_U \frac{|U^T S_b U|}{|U^T S_w U|} \tag{2}$$

where $|\cdot|$ is the matrix determinant operator. The reason to use determinant ratio is that, the scattering distribution of the samples is directly proportional to the determinant of the scatter matrix (Duda et al., 2000; Fukunaga, 1991). As another word, the absolute value of the determinant of a group of real-valued vectors is equal to the volume of the parallelepiped spanned by these vectors (Strang, 1993). Before introducing the way to optimize Eq. (2), we firstly re-visit the following lemma (Zhou and Huang, 2001):

Lemma 1. ***Scatter Ratio Invariance****: If the sample data points are multiplied by any invertible $d\times d$ matrix R, the scatter ratio $J(U)$ remains the same.*

The proof is very trivial, as the determinants of R on the numerator and the denominator canceled out.

Due to exactly the same reason, if the projection matrix U is multiplied by any invertible $r\times r$ matrix V, $J(U)$ does not change either. Based on this observation, we can easily conclude the following lemma:

Lemma 2. ***Solution Ambiguity****: If $U*$ is the solution of Eq. (2), then $VU*$ is also the solution of Eq. (2), for any invertible $r\times r$ matrix V.*

Lemma 2. *suggests that, it is safe to fix $U^T S_w U = I$ in Eq. (2) without any possibility to lose the optimal solution, because if $U*^T S_w U* = ZZ^T$ [1], then there must be another $U*' = U*(Z^T)^{-1}$ such that $U*'^T S_w U*' = I$. Therefore Eq. (2) becomes:*

$$J(U) = \max_{U^T S_w U = I} |U^T S_b U| \tag{3}$$

where the solution is obviously the Generalized Eigen-Value

* Corresponding author.
E-mail address: ypguan@shu.edu.cn (Y. Guan).

[1] For mathematical rigor, the operator ^{T} should be referred to as the matrix conjugate transpose operator, even this paper only discusses real numbers. Because

http://dx.doi.org/10.1016/j.engappai.2015.03.006
0952-1976/

Decomposition (GEVD) (Duda et al., 2000; Fukunaga, 1991) of $(S_w)^{-1}S_b$.

Thanks to Lemma 2, the solution of Eq. (2) becomes straight forward; likewise, owing to Lemma 2, the solution of Eq. (2) can fill up the entire subspace spanned by any $U*$, and thus making it ubiquitous. For the sake of making the resulted DR projection not so arbitrary, there are various streams of LDA research directions, *e.g.* the orthogonal LDAs (Ye and Xiong, 2006; Nie et al., 2009) impose the additional constraint $U^TU=I$.

Actually, the determinant is a very 'rough' measure for matrices. One particular fact accounting for such argument is that, the determinant of any orthogonal matrix is $+1$ or -1, regardless of how the orthogonal transform is. It is not a nice property to measure the DR projections used in statistical applications.

But the motivation of previous LDA literature to replace the determinant ratio with the trace ratio is from another perspective: it is argued that the trace ratio is a natural criterion in discriminant analysis as it directly connects to the Euclidean distances between training data points (Nie et al., 2009; Wang et al., 2007). Besides the determinant ratio problem is equivalent to the ratio trace problem, which is an inexact approximation of the trace ratio problem (Fukunaga, 1991; Wang et al., 2007). Furthermore, by comparing Eqs. (1) and (2), it is not difficult to understand that, the determinant ratio is not the original Fisher's criterion (Fisher, 1936).

Using the determinant ratio is just one somewhat inappropriate way to extend the original Fisher's LDA (Fisher, 1936). Another inappropriate but more implicit extension is about how to generalize it for discriminating multiple classes, also known as Multiple Discriminant Analysis (MDA) (Duda et al., 2000; Rao, 1948). About a decade ago, the non-optimality of Eq. (2) with respect to minimizing the classification error rate in the low-dimensional subspace was realized (Loog et al., 2001), because outlier classes dominate the eigenvalues decomposition, and the MDA tends to over-weight the influence of classes that are already very well-discriminated (see Fig. 1(a)). In this paper, we generalize the original Fisher's LDA by projecting data points into multiple subspaces, which is mathematically reasonable and fundamental, as if only two classes are needed to be discriminated in each subspace (see Fig. 1(b)). We adapt this idea into two existing LDA methods, and find that multi-subspace DR can balance the recognition rates of well-discriminated (outlier) classes and poorly-classified (similar) classes.

In recent years, the LDA continues to be one of the most active topic for a variety of research fields, such as pattern recognition, computer vision, and artificial intelligence. Chen et al. (2013) developed the Complete Large Margin Linear Discriminant Analysis (CLMLDA) that constructs two mathematical programming models by maximizing the minimum distance between each class center and the total class center respectively in the null space of within-class scatter matrix and its orthogonal complementary space. Shao et al. (2011) proposed the Sparse Linear Discriminant Analysis (SLDA) method, to handle the situation where the number of dimension d is much larger than the training sample size N, and the covariance matrix satisfies some sparsity conditions. SLDA solves a specific type of LDA, and it still considers the two-class classification problem, while our proposed methods are designed to handle the classification problem with a large number of class.

The rest of the manuscript is organized as follows: Section 2 formulates a single-label classification problem and presents the proposed methods; Section 3 justifies and analyzes why our methods work and their advantages; Experimental results are discussed in Section 4 and finally, Section 5 draws concluding remarks and points out some possible future work.

2. Problem formulation and our methods

In this section, the proposed definition of multi-subspace scatter matrices are firstly presented, based on which, two popular objective functions are adopted, leading to the multi-subspace ratio-trace LDA (ms-LDA) and trace-ratio LDA (ms-LDA-tr) methods. Finally, we discuss how to combine the multiple subspaces by a simple k-Nearest Neighour (kNN) classifier.

2.1. Definition of multi-subspace scatter matrices

Given a dataset of N samples, denote $x_i \in \mathbb{R}^d$ $(1 \le i \le N)$ as its feature vectors, and y_i $(1 \le i \le N)$ as its corresponding labels. As a single-label classification problem with K classes, $y_i \in \mathbf{Z}$, $1 \le y_i \le K$. Let the input N samples be partitioned into K groups as π_k $(1 \le k \le K)$, where π_k denotes the sample set of the k-th class with $|\pi_k|$ data points.

In the traditional LDA, S_b and S_w are defined as:

$$S_b = \sum_{k=1}^{K} |\pi_k| (m_k - m)(m_k - m)^T \tag{4}$$

$$S_w = \sum_{k=1}^{K} \sum_{x_i \in \pi_k} (x_i - m_k)(x_i - m_k)^T \tag{5}$$

where m_k is the mean (*i.e.* class centroid) of the k-th class, m is the mean of all N data points (*i.e.* global centroid). Typically $K \ll N$. The rank of S_b and S_w are at most $K-1$ and $N-K$ respectively, making the rank of $(S_w)^{-1}S_b$ at most $K-1$. That is the reason why at most $r=K-1$ projection directions can be obtained from the traditional LDA. It seems that, computing S_b only with the class centroid loses lots of useful information[2] and makes the rank of S_b too low. So a question arises naturally: why not use the individual data points directly (as shown below)?

$$S_t = \sum_{i=1}^{N} (x_i - m)(x_i - m)^T. \tag{6}$$

Here S_t turns out to be the covariance matrix of X. Interested readers can verify that $S_t = S_b + S_w$. Due to this reason, S_t is also called the total scatter matrix. Replacing S_b with S_t in Eq. (2) leads to:

$$J(U) = \max_U \frac{|U^T S_t U|}{|U^T S_w U|}, \tag{7}$$

where the solution, the GEVD of $(S_w)^{-1}S_t$, has exactly the same eigenvectors with that of $(S_w)^{-1}S_b$, and all eigenvalues added by 1. Therefore, the latter approach to construct S_b, namely the individual-based S_b, brings some useless computation with results identical to the traditional centroid-based S_b.

But such situation is not the same in their multi-subspace versions. Following our basic class-adaptive idea, each class π_k has its own S_b^k and S_w^k for deriving its DR projection separately. The individual-based S_b^k represents the sum of the distances between all individual data points ***not*** belonging to the k-th class π_k and the

(*footnote continued*)
$U*S_wU*^T$ is real-valued, symmetric and positive semi-definite, it can always be decomposed into ZZ^T by the Cholesky decomposition (Duda et al., 2000).

[2] Though it is believed theoretically that, no information is lost if all data points are assumed to be normally distributed in each class and $r \ge K-1$ (Loog et al., 2001).