Neurocomputing 73 (2010) 1808–1817

Contents lists available at ScienceDirect

Neurocomputing

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

Quotient vs. difference: Comparison between the two discriminant criteria

Yuting Tao, Jian Yang *

School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, PR China

ARTICLE INFO

Available online 19 March 2010

Keywords:
Feature extraction
Fisher linear discriminant analysis
Quotient criterion
Difference criterion
Small sample size problem
Large sample size problem

ABSTRACT

Two discriminant criteria—quotient and difference, are commonly used in linear discriminant analysis. In the paper, we experiment with the CENPARMI handwritten numeral database, the NUST603 handwritten Chinese character database, the ORL face image database and the FERET face image database and find that the quotient criterion is better than the difference criterion for large sample size problems such as the character recognition, while the difference criterion is better for small sample size problems such as face recognition. Through theoretical analysis, the defect of the difference criterion—the correlation among discriminant vectors is revealed, and it is testified that the quotient criterion is superior to the difference criterion in general, if the instability of denominator can be overcome. Otherwise, the difference criterion might be better. Finally, the two criteria (quotient and difference) are unified into one framework in the paper.

1. Introduction

In the past years, many works have been done to deal with the issue of classification [12]. In summary, classification can be categorized into two parts in general—supervised and unsupervised. The former focuses on the known class label in the training samples, the most widely used method is Fisher linear discriminant analysis (FLDA) [23,7,4]; while the latter stresses on using the unknown class label samples to evaluate what class they belong to, for which several methods can be taken, such as locality preserving projection (LPP) and so forth [10,14]. And in this paper, the statement and analysis on the supervised FLDA method is provided.

The classical LDA is raised by Fisher in 1936 for the first time, aiming at seeking some directions for projection, onto which the within-class scatter should be minimized while the between-class scatter should be maximized [23]. In 1975, Foley and Sammon did some improvement work for FLDA, and pointed out that the discriminant vectors should be orthogonal mutually [24]. The famous Rayleigh quotient is successfully applied to FLDA by maximizing the ratio of the between-class scatter to the within-class scatter if projected onto some directions, which is called quotient criterion. And it is quite widely used in linear discriminant analysis [1,4,5,7,19,20,27].

Another one is difference criterion, in which the projection axis are attempted to be found where the result after between-class scatter deducting within-class scatter should be as large as possible [2,3,13,17,28]. Both of the two criteria are consistent with the main idea of maximal between-class information, as well as minimal within-class information [2,3,16].

LDA can be performed directly in the original input space with respect to large sample size cases, such as the character recognition, because the number of training sample surpasses the dimension of the original input vector space, thus the problem of singularity can be avoided. However, with respect to small sample size problems such as face recognition, doing LDA transformation directly will encounter with singularity and computational difficulty. PCA transformation is operated in order to bring about convenience to LDA. For the sake of expressing and retaining the original useful information in high dimension and mapping it into low dimension as much as possible, it can be solved conveniently in PCA transformation by singular value decomposition (SVD) [5]. After that, the vector space has been reduced from high to low, and the minimum square error is ensured when reconstruction work is done [21,22]. It has been evidenced that, doing PCA transformation in the first step, and doing LDA transformation in the second step is equal to doing LDA directly in the high dimensional original input space [1]. It can be concluded that, PCA plus LDA is a common combined method in dealing with feature extraction and classification towards high dimensional and small sample size cases [6,11,15].

In this paper, we systematically compare the two criteria—quotient and difference, from the viewpoints of stability and correlation respectively. Although singularity of total-scatter can be eliminated in PCA transformed space, the within-class scatter located in the denominator of Rayleigh quotient is still instable because of the small sample size problem [18]. However, the problem does not exist in difference criterion, since the stability of the within-class scatter is taken into no consideration. So it is

* Corresponding author.
E-mail addresses: tao.yuting@yahoo.com.cn (Y. Tao), csjyang@mail.njust.edu.cn (J. Yang).

0925-2312/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2009.10.026

often seen that difference criterion can demonstrate better recognition accuracy, compared to quotient criterion, in terms of small sample size.

As for large sample size, the within-class scatter in the denominator is stable, so quotient criterion is superior to difference criterion generally. This is because the former can generate uncorrelated optimal discriminant vectors in LDA transformation while the latter cannot. That is to say, the discriminant vectors produced by difference criterion are not orthogonal mutually, which put adverse influence in recognition.

Fortunately, regularization can be done to make the within-class scatter in the denominator a little bigger, to overcome the instability in quotient. Hence, the recognition rate can improve to be much better [4]. Finally, the two criteria might be unified into one framework, to forge one function to do LDA. After that, the recognition rate can arrive at the peak, when the parameter in the function can be varied within some certain scope, to adjust the balance between quotient and difference.

The following sections in the paper is arranged like this: some introduction of FLDA is presented in Sections 2 and 3 shows the experiments and the results on large and small sample size cases, analysis for the experimental results is given in Section 4, the combination of the two criteria is discussed in Section 5, and a final conclusion is made in Section 6.

2. Linear discriminant analysis (LDA)

Suppose that L is the number of classes, and N_c is the number of samples in class c, so the equation $\mu_c = 1/N_c \sum_{i \in c} x_i$ represents the mean of samples in class c. Suppose that N is the total number of all samples, then $\bar{x} = 1/N \sum_i x_i = 1/N \sum_c N_c \mu_c$, and this equation stands for the total mean of the whole samples.

The between-class scatter matrix S_b, the within-class scatter matrix S_w, and the total-scatter scatter matrix S_t can be respectively defined as follows [7]

$$S_b = \sum_c N_c(\mu_c - \bar{x})(\mu_c - \bar{x})^T \quad (1)$$

$$S_w = \sum_c \sum_{i \in c} (x_i - \mu_c)(x_i - \mu_c)^T \quad (2)$$

$$S_t = \sum_i (x_i - \bar{x})(x_i - \bar{x})^T \quad (3)$$

And it is easy to show that $S_t = S_b + S_w$.

2.1. Two discriminant criteria based LDA methods

Quotient-LDA

The quotient criterion (i.e. Fisher criterion) is defined as follows

$$J(W) = \frac{W^T S_b W}{W^T S_w W} \quad (4)$$

Maximizing the quotient criterion can be converted into finding the generalized eigenvectors of $S_b W = \lambda S_w W$, or equivalently finding the eigenvectors of $S_w^{-1} S_b$. Since $S_t = S_b + S_w$, the quotient criterion can be rewritten as

$$J(W) = \frac{W^T S_b W}{W^T S_t W} \quad (5)$$

So the problem is equivalent to find the generalized eigenvectors of $S_b W = \lambda S_t W$.

Difference-LDA

The difference criterion is defined by

$$J(W) = W^T (S_b - \Delta \cdot S_w) W, \quad \text{subject to } W^T W = 1, \quad (6)$$

where Δ is the parameter to adjust the magnitude of S_w to balance S_b and S_w.

Maximizing the difference criterion can be converted into finding the eigenvectors of $(S_b - \Delta \cdot S_w)$. Suppose that $\eta_1, \eta_2 \ldots\ldots \eta_m$ are the eigenvectors of $(S_b - \Delta \cdot S_w)$ corresponding to the m largest eigenvalues. Then, the transformation matrix of Difference-LDA is $[P = [\eta_1, \eta_2 \ldots\ldots \eta_m]$.

If S_w is singular, with the increase of the value of Δ from 0 to infinite, the discriminant vectors of difference criterion are more and more from the null space of S_w. When Δ approaches infinite, the discriminant vectors are solely from the null space of S_w because $\lim_{\Delta \to \infty} W^T S_w W = 0$. For more details, refer to [2,17].

If S_w is nonsingular, with the increase of Δ, the largest eigenvalue from Eq. (6) decreases continuously, even to be negative (i.e.$S_b < \Delta \cdot S_w$, if projected onto any direction). It leads to much lower recognition rate, because the between-class information and the within-class information overlap together [17]. Therefore, the value of Δ should not be too large.

2.2. Distinction of two methods from the correlation point of view

In this subsection, we will show the Quotient-LDA transformed features are mutually uncorrelated, while the Difference-LDA transformed features not.

Theorem. (Yang et al. [8]) *There exists n eigenvectors $\varphi_1, \varphi_2, \ldots, \varphi_n$ of $S_b W = \lambda S_t W$ which satisfy the following property*:

$$\varphi_i^T S_t \varphi_j = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad i,j = 1, \ldots, n, \quad (7)$$

$$\varphi_i^T S_b \varphi_j = \begin{cases} \lambda_i & i = j \\ 0 & i \neq j \end{cases} \quad i,j = 1, \ldots, n, \quad (8)$$

where $\lambda_i (i = 1, \ldots, n)$ is the generalized eigenvalue corresponding to the eigenvector φ_i $(i = 1, \ldots, n)$.

If we suppose $\Phi = [\varphi_1, \varphi_2, \ldots \varphi_d]$ is a set of optimal discriminant vectors corresponding to the largest d positive eigenvalues, and Y is the original set of vectors, then after the Quotient-LDA transformation, we get a new transformed d-dimensional space: $Z = \Phi^T Y$, and of which $Z_i = \varphi_i^T Y (i = 1, 2, \ldots d)$

The covariance between Z_i and Z_j is:

$$\mathrm{Cov}(Z_i, Z_j) = E(Z_i - EZ_i)(Z_j - EZ_j) = \varphi_i^T \{E(Y - EY)(Y - EY)^T\} \varphi_j = \varphi_i^T S_t \varphi_j \quad (9)$$

Accordingly, the correlation coefficient between Z_i and Z_j is defined as

$$\rho(Z_i, Z_j) = \frac{\varphi_i^T S_t \varphi_j}{\sqrt{\varphi_i^T S_t \varphi_i}\sqrt{\varphi_j^T S_t \varphi_j}} \quad (10)$$

Therefore, if $\rho(Z_i, Z_j) = 0$, and $i \neq j$, then Z_i and Z_j is uncorrelated mutually. In a word, the Quotient-LDA transformation can eliminate the correlations between features [9,25,26].

The Difference-LDA, however, does not hold this property. This is because that the discriminant vectors of Difference-LDA (i.e. the eigenvectors of the matrix $(S_b - \Delta \cdot S_w)$) are orthogonal, rather than S_t-orthogonal.

3. Experiments and analysis

In this section, the performance of Difference-LDA and Quotient-LDA is done on the CENPARMI handwritten numeral database, the NUST603 handwritten Chinese character database, the ORL face image database and the FERET face image database.