

Automatic Coefficient Selection in Weighted Maximum Margin Criterion*

Zhengdong Cheng^{1,2}, Bin Shen¹, Xiang Fan^{2,3}, and Yu-Jin Zhang¹

1. Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China

2. Electronic Engineering Institute, Hefei, 230037, China

3. University of Science and Technology of China, Hefei 230027, China

{czd06, chenbin03}@mails.tsinghua.edu.cn;

FanXiangLXL@163.com; zhangyj@ee.tsinghua.edu.cn

Abstract

In this paper, one of the problems of Linear Discriminant Analysis (LDA), that is, it pays more attention on minimizing the within-class scatter than on maximizing the between-class scatter, is treated. Though the Weighted Maximum Margin Criterion (WMMC) with an appropriate weighted coefficient can solve this problem, how to select this coefficient automatically is still difficult as most of previous works determine it manually. To deal with this problem, a novel approach to determine the coefficient automatically is proposed. The description and analysis of the approach are given in details, and the experiments on BernDBS face database and JAFFE expression database are performed. The results show that the WMMC with the weighted coefficients determined by this novel approach outperforms traditional LDA.

1. Introduction

In dealing with pattern classification problems, Linear Discriminant Analysis (LDA) is a successful dimension reduction method when we face with high dimensional data. LDA aims to find the optimal project direction so that the between-class scatter is maximized while the within-class scatter is minimized. As we know, LDA has one typical drawback that it may suffer from the small sample size (SSS) problem in dealing with high dimensional data [1]. The SSS problem leads the within-class scatter matrix to be singular, which makes LDA difficult to work. Many techniques have been proposed to deal with this problem, such as RLDA [2], NLDA [3, 4], etc.

However, LDA has another latent drawback in it. In fact, we cannot maximize the between-class scatter while minimizing the within-class scatter at the same

time. We can only make a tradeoff between them. LDA is one method which is aimed to make this tradeoff. But the criterion function of LDA is the ratio of the between-class scatter to the within-class scatter. When it is being maximized, minimizing the denominator, i.e. the within-class scatter, always receives more attention than maximizing the numerator, i.e. the between-class scatter. So, LDA cannot make a good enough tradeoff.

Fortunately, the Weighted Maximum Margin Criterion (WMMC) [5] provides us a way to overcome this drawback and make a better tradeoff. However the performance of WMMC is directly dependent on its weighted coefficient, and how to determine a proper one is not an easy task. Many previous methods to determine this coefficient are usually time consuming exhaustive ones. In this paper, we will propose a novel method that is able to automatically select a proper weighted coefficient in WMMC. Our experimental results show that WMMC with the coefficient selected by our algorithm outperforms the traditional LDA.

The rest of this paper is organized as follows: In section 2 we briefly review the related work on LDA and WMMC; in section 3, we present the novel method to select the weighting coefficient; the experimental results are listed in section 4 with some discussions, and we conclude this paper in section 5.

2. Related Works

Here we will discuss a pattern classification problem in which the dimension of the original sample space is d , and the number of the samples is n .

Let x_i^j be the j -th sample data in class i , where $j = 1, 2, \dots, n_i$, $i = 1, 2, \dots, c$, and n_i is the number of sample in class i . Then the between-class scatter matrix and the

* This work has been supported by Grants NNSF-60573148.

within-class scatter matrix are defined as

$$S_w = \sum_{i=1}^c \sum_{j=1}^{n_i} (x_i^j - m_i)(x_i^j - m_i)^T \quad (1)$$

$$S_b = \sum_{i=1}^c n_i (m_i - m)(m_i - m)^T \quad (2)$$

where m_i ($i = 1, 2, \dots, c$) is the mean of the data in class i , and m is the mean of all the data:

$$m_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_i^j \quad (3)$$

$$m = \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^{n_i} x_i^j \quad (4)$$

2.1 LDA and its drawback

LDA is to learn a discriminant projection matrix $W \in \mathbf{R}^{d \times m}$. Then the high-dimensional data x is projected into a m -dimensional subspace by

$$y = W^T x \quad (5)$$

To achieve this W , LDA maximizes the following criterion function [6]:

$$J_F(W) = \frac{\text{tr}(W^T S_b W)}{\text{tr}(W^T S_w W)} \quad (6)$$

However LDA has a latent drawback in it. To be easy to understand, we illustrate an example firstly. Let $a_1 = 3$, $a_2 = 2$, $b = 1$, it is easy to know that

$$\frac{a_1}{a_2} < \frac{a_1 + b}{a_2} < \frac{a_1}{a_2 - b} \quad (7)$$

This example shows that, if one wants to increase the ratio, minishing the numerator is more efficient than increasing the denominator. When maximizing the ratio $J_F(W)$, it may decrease the denominator (*i.e.* the within-class scatter) more. So LDA may not get a good enough tradeoff between the maximization of between-class scatter and the minimization of within-class scatter.

2.2 WMMC

The goal of WMMC is the same with LDA, but it achieves its purpose by maximizing the criterion function below [5]:

$$J_M(W) = \text{tr}\{W^T (S_b - \beta S_w) W\} \quad (8)$$

where β is the weighted coefficient.

Obviously, maximizing $J_M(W)$ is to maximize the between-class scatter when $\beta = 0$, and is to minimize the within-class scatter when $\beta = \infty$. So there must be a

proper weighted coefficient $\beta \in (0, \infty)$, which may make WMMC balance in maximizing the between-class scatter and minimizing the within-class scatter.

However, it is still an open problem to find a proper weighted coefficient. Most of existing works which employ WMMC usually have the weighted coefficient selected manually, such as Liu [7], Wang [8] and Wang [9]. In the following section, we will propose a method to determine it, which is performed automatically.

3. Our Work

In this section, we present the idea and details of our work.

Obviously, the proper weighted coefficient should depend on the between-class scatter matrix S_b and the within-class scatter matrix S_w . One method for selecting the coefficient is

$$\beta_1 = \frac{\text{tr}(S_b)}{\text{tr}(S_w)} \quad (9)$$

But this coefficient β_1 neglects the fact that it is in m -dimensional space, not in d -dimensional original space, that we maximize the between-class scatter while minimizing the within-class scatter. So the proper weighted coefficient in WMMC should also be dependent on the relation between the between-class scatter matrix and the within-class matrix in m -dimensional space, *i.e.* $W^T S_b W$ and $W^T S_w W$, and can be formed as

$$\beta = \frac{\text{tr}(W^T S_b W)}{\text{tr}(W^T S_w W)} \quad (10)$$

In equation (10), we need to first determine a matrix $W \in \mathbf{R}^{d \times m}$ which should make the difference of its effect on S_b and S_w small. We know that the $W^T S_b W$ is as significant as $W^T S_w W$ in the function

$$\begin{aligned} J_1(W) &= \text{tr}\{W^T (S_b - S_w) W\} \\ &= \text{tr}(W^T S_b W) - \text{tr}(W^T S_w W) \end{aligned} \quad (11)$$

Maximizing $J_1(W)$, we can get a column orthogonal matrix $Q \in \mathbf{R}^{d \times m}$ whose columns are the m leading eigenvectors of matrix $(S_b - S_w)$. If we select Q as W in equation (10), then the ratio

$$\beta_2 = \frac{\text{tr}(Q^T S_b Q)}{\text{tr}(Q^T S_w Q)} \quad (12)$$

may encode the information of the relation between $W^T S_b W$ and $W^T S_w W$.

It is intuitive that the proper weighted coefficient should contain not only the information about the relation between S_b and S_w in original space but also

that about the relation between $\mathbf{Q}^T \mathbf{S}_b \mathbf{Q}$ and $\mathbf{Q}^T \mathbf{S}_w \mathbf{Q}$ in the reduced dimensional space. Based on the discussion above, now we finally propose our selected weighted coefficient as

$$\beta_T = \frac{\text{tr}(\mathbf{Q}^T \mathbf{S}_b \mathbf{Q})}{\text{tr}(\mathbf{Q}^T \mathbf{S}_w \mathbf{Q})} + \frac{\text{tr}(\mathbf{S}_b)}{\text{tr}(\mathbf{S}_w)} \quad (13)$$

where \mathbf{Q} is same as in equation (12).

Furthermore, we know that the trace of a matrix is the sum of the diagonal components of the matrix, while its Frobenius norm is square root of the sum of the square of all the matrix components. So the Frobenius norm deploys more information of the matrix than its trace. Therefore we can also select the weighted coefficient in WMMC as

$$\beta_F = \frac{\|\mathbf{Q}^T \mathbf{S}_b \mathbf{Q}\|}{\|\mathbf{Q}^T \mathbf{S}_w \mathbf{Q}\|} + \frac{\|\mathbf{S}_b\|}{\|\mathbf{S}_w\|} \quad (14)$$

where $\|\cdot\|$ is Frobenius norm of a matrix. It should probably be more efficient than β_T , which is later tested by our experiments.

As can be seen from the above, our method can be simply described as follows:

(i) Compute the m leading eigenvectors of $(\mathbf{S}_b - \mathbf{S}_w)$ to form the matrix \mathbf{Q} which is a column orthogonal matrix.

(ii) Select β_T in (13) or β_F in (14) as the weighted coefficient of WMMC.

4. Experiments

We carry out two experiments to verify the efficiency of our method on the BernDBS face database and the JAFFE expression database.

To compare our method with NLDA, we first reduce the original images by Principal Component Analysis (PCA) to $n-1$ dimension, where n is the number of the training samples. We do not compare our method with other methods proposed in [10-12], because they have same accuracy as NLDA when the dimension of the null space of the within-class scatter matrix is equal to $c-1$. We adopt the nearest neighbor classifier with 2-norm.

4.1 The BernDBS face database

The BernDBS face database is made up of 30 persons. Each person has 10 images. There are some variations between different images, such as light condition, expression, accessory, etc.

We randomly select k ($=2, 3, 4, 5$) images of each person for training and the rest $10-k$ images of each

person are left for testing. 10 rounds are performed for each value of k , and Table 1 shows the average recognition accuracy (%). Here no any pre-processing has been done.

Table 1. Recognition Accuracy on BernDBS database

method	$k=2$	$k=3$	$k=4$	$k=5$
NLDA	69.04	82.00	88.33	90.93
WMMC- β_1	65.83	77.90	83.39	86.87
WMMC- β_2	69.21	82.81	88.72	92.07
WMMC- β_T	69.42	82.86	89.11	92.53
WMMC- β_F	69.42	83.05	89.44	92.40
Best sample (β)	69.54 (25)	83.48 (32)	89.61 (18)	92.53 (12)

To verify the efficiency of our proposed algorithm, we sample the β from 1 to 50, at step of 1, i.e. $\beta = 1, 2, \dots, 50$. For each value of β , we test the accuracy of WMMC with β as the weighted coefficient, and show the best accuracy of these 50 experiments in the last row of the table. At same time, we also list the value of β with best accuracy in the brackets.

Moreover, take $k = 5$ as an example, Figure 1 illustrates the recognition accuracy with weighted coefficient in WMMC ranging from 1 to 50. The computed β_T and β_F are at the places where the accuracy is high enough.

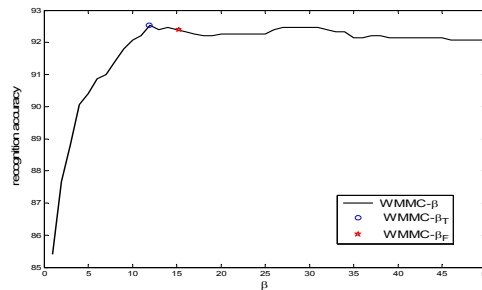


Figure 1 Recognition Accuracy with β (ORL)

4.2 The JAFFE Expression Database

The JAFFE face expression database contains 213 images posed by ten female Japanese. Every person has several images for each of the seven expressions: anger, disgust, fear, happy, neutral, sad, and surprise. In the experiment, every original image has been cropped manually into smaller image of the size 128×128 pixels. We remove the two images named "KR.SR3.79.tiff" and "NA.SU1.79.tiff", for their labels seem to be wrong.

We adopt the “leave-one-person-out” cross validation strategy. That is, the images of seven expressions belong to one person are selected as testing set and the rest images are used as the training set. We repeat this for 10 turns until every person is used in the testing set. The final recognition accuracy is the average of all the recognition results. Table 2 shows the final recognition accuracy (%).

Table 2. Recognition Accuracy on JAFFE database

method	accuracy	method	accuracy
NLDA	60.09	WMMC- β_T	65.32
WMMC- β_1	42.45	WMMC- β_F	65.32
WMMC- β_2	65.32	sample	65.77(3)

Again, we sample β from 1 to 50, at step of 1. The best result of the fifty experiments is also shown in the last row of Table 2. Figure 2 gives the average accuracy with the weighted coefficient changing from 1 to 50 in WMMC. This figure still reveals that β_T and β_F are near the optimal choice.

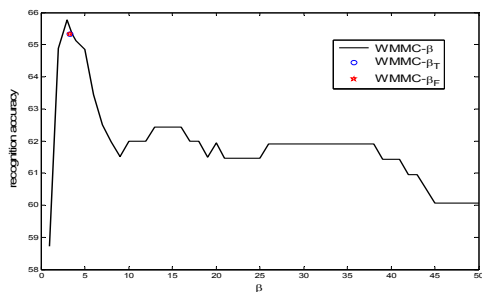


Figure 2 Recognition Accuracy with β (JAFFE)

4.3 Discussions

Based on the above tests on two databases, two more points are discussed here.

When the dimension of the null space of the within-class scatter matrix is equal to $c-1$, maximizing the $J_F(\mathbf{W})$ is equal to maximizing the $J_M(\mathbf{W})$ with $\beta = \infty$ [6], while the WMMC with $\beta = \infty$ is equal to NLDA [7]. So we can see that the NLDA pays too much attention to minimize the within-class scatter. This can be verified by our experiments.

WMMC with a proper coefficient can overcome the drawback of LDA. However, if the coefficient is not properly chosen, the WMMC will perform poorly, such as β_1 , i.e. the performance of WMMC is sensitive to the value of weighted coefficient. From the experimental results, we can see that our proposed weighted coefficient works effectively.

5. Conclusions

In this paper, we point out that, to find the optimal project direction, we must balance maximizing the between-class scatter and minimizing the within-class scatter. To automatically make a satisfying tradeoff between them, we propose two ways to select the weighted coefficient for WMMC based on the information of \mathcal{S}_b and \mathcal{S}_w . The effectiveness of our method is confirmed by the experiments on face database and expression database.

References

- [1] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 1990.
- [2] Ziquan Hong and J.Y. Yang, Optimal discriminant plane for a small number of samples and design method of classifier on the plane, *Pattern Recognition*, 24(4), 317–324, 1991.
- [3] Lifan Chen, H.Y. M. Liao, M.T. Ko, J.C. Lin, and G.J. Yu, A new LDA-based face recognition system which can solve the small sample size problem, *Pattern Recognition*, 33, 1713–1726, 2000.
- [4] Yuefei Guo, L.D. Wu, H. Lu, Z. Feng, X.Y. Xue, Null Foley–Sammon transform, *Pattern Recognition*, 39, 2248–2251, 2006.
- [5] Wenming Zheng, C.R. Zou and L. Zhao, Weighted maximum margin discriminant analysis with kernels, *Neurocomputing*, 67, 357–362, 2005.
- [6] Huan Wang, S.C. Yan, D.Xu, X.O. Tang, T. Huang, Trace Ratio vs. Ratio Trace for Dimensionality Reduction, *IEEE Conference on Computer Vision and Pattern Recognition*, 17–22, 2007.
- [7] Jun Liu, S.C. Chen and X.Y. Tan, A study on three linear discriminant analysis based methods in small sample size problem, *Pattern Recognition*, 41(1), 102–116, 2008.
- [8] Haixian Wang, W.M. Zheng, Z.L. Hu, Local and Weighted Maximum Margin Discriminant Analysis, *IEEE Conference on Computer Vision and Pattern Recognition*, 17–22, 2007.
- [9] Haixian Wang, S.S. Chen, Z.L. Hu, Image Recognition Using Weighted Two-Dimensional Maximum Margin Criterion, *Third International Conference on Natural Computation*, 1(24), 582–586, 2007.
- [10] Jian Yang, J.Y. Yang, An Optimal FLD algorithm for facial feature extraction, *SPIE Proceedings of the Intelligent Robots and Computer Vision XX: Algorithms, Techniques, and Active Vision*, 4572, 438–444, 2001.
- [11] Fengxi Song, D. Zhang, J.Y. Yang, A novel dimensionality-reduction approach for face recognition, *Neurocomputing Letters*, 69, 1683–1687, 2006.
- [12] A.K. Qin, P.N. Suganthan, M. Loog, Generalized null space uncorrelated Fisher discriminant analysis for linear dimensionality reduction, *Pattern Recognition*, 39, 1805–1808, 2006.