MARGIN-MAXIMIZATION DISCRIMINANT ANALYSIS FOR FACE RECOGNITION

Yan Zhu, Eric Sung

Division of Control and Instrumentation, School of EEE
Nanyang Technological University, Nanyang Ave, Singapore 639798

ABSTRACT

LDA and its variants are popular for image-based classification problems such as face recognition. However, their performance is inherently unstable when the samples are sparse. In this paper, we propose a new type of discriminant analysis called MMDA, which derives features by maximizing the average margin between the classes. The method does not require S_W to be non-singular and well-conditioned as it does not involve its inverse term, and the features can be directly derived from the input space. A computational trick has also been proposed for MMDA to handle high-dimensional data. We conduct intensive tests on ORL and UMIST face databases, and the results show that MMDA is a good replacement of LDA for sparse sample problem.

1. INTRODUCTION

Face recognition is a typical task of image-based pattern classification. For such application, we often face the sparse sample problem, where the dimension of the input data is much higher than the number of the training samples. Many techniques have been developed to reduce the data dimension and at the same time extract the useful features for classification. Among them, Linear Discriminant Analysis (LDA) and its linear/non-linear variants have been shown to be very successful for multi-class problems [1] [2] [3] [4].

LDA, also known as FDA, finds features as projection directions that maximize the Fisher's discriminant $\mathcal{J}_F(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}$ [1]. S_B and S_W are between- and within-class scatter matrices as obtained from the training samples. It can be seen that $\mathcal{J}_F$ actually measures a type of normalized distance between the class centers. However, the normalization term S_W^{-1} not only causes computational inefficiency, but also makes LDA inherently unstable.

Extra concerns are thus required to handle the singular or close to singular S_W, which is almost always the case for sparse sample problems. Tian et al. (1986) [5] use the pseudo-inverse to compute S_W^{-1}. In [6], small perturbation is added to S_W to make it nonsingular. Other than the computational approach, Belhumeur et al. [1] and Yang et al. [2] avoid the singular S_W by discarding the null-space of total or between-class scatter matrices. Also, some regularization schemes have been proposed, which apply parametric shrinkage to individual class scatters, and in effect, help to bias the feature search away from null space (e.g. [7] and [4]). But all these methods do not essentially solve the problem, and some may greatly complicate the system.

Now, the question is whether the S_W^{-1} term is necessary in the formulation of a class-discriminant analysis. For classification tasks, we actually prefer features with least overlap among the classes. This can be measured as the average margin between the class edges. In this paper, we will show that the average margin between the classes can be estimated in terms of $S_B - \beta S_W$ (β is a constant), and we propose our Margin-Maximization Discriminant Analysis (MMDA) based on this. MMDA avoids the inverse of S_W, and hence shall be more stable for the sparse sample problem. Here, β is a parameter that quantifies the within-class spread. Note that if $\beta = -1$, MMDA is equivalent to Principle Component Analysis (PCA).

In a recent paper of Li et al. [8], a method called Margin Maximization Criterion (MMC) has also been proposed, which is equivalent to our MMDA at $\beta = 1$. But in our method, the margin measure is actually used to reflect the non-overlap region between the classes, and we will show later that $\beta \gg 1$, e.g. $\beta \approx 9$, is a preferred setting. In fact, β can be viewed as a regularization parameter, because increasing β value actually reduces the feature searching space. But since it does not change the margin measure essentially, we expect the MMDA performance to be stable at a proper β setting.

MMDA can also be formulated in the non-linear space using the kernel tricks (c.f. [8]). But in this paper, we focus on its linear version to see the fundamental difference between MMDA and LDA. To handle the high dimensional data, a computational trick of MMDA is also proposed. We compare MMDA to a popular LDA implementation as in [1]. The feature size retained after the PCA dimension reduction step, K_{pca}, has been varied to see how unstable the LDA is. In the rest of the paper, we first show the derivation of the MMDA method. The computational trick is then proposed. The performance of MMDA is evaluated and compared to LDA on ORL and UMIST face databases. Finally, we conclude our findings.

0-7803-8554-3/04/$20.00 ©2004 IEEE.

2. THE MARGIN-MAXIMIZATION DISCRIMINANT ANALYSIS (MMDA)

We first derive the linear version of the new type of discriminant analysis. This clearly shows its underlying principle and hints at the selection of the β value. The nonlinear version of MMDA can be similarly derived using the kernel trick.

Suppose we have N d-dimensional sample patterns $\{\mathbf{x}_k\}$ belonging to C different classes, and assume that there are N_i samples for class i that is defined by subset $\mathcal{C}_i$. The Between-class and Within-class Scatter can be given in the expectation form:

$$S_B = E_i\{(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T\}, \quad \text{and} \quad (1)$$

$$S_W = E_i\{E_k\{(\mathbf{x}_k - \mathbf{m}_i)(\mathbf{x}_k - \mathbf{m}_i)^T\}\} \quad (\mathbf{x}_k \in \mathcal{C}_i) \quad (2)$$

where $\mathbf{m}$ is the global mean, $\mathbf{m}_i$ is the class mean, and $E_i\{\cdot\}$ denotes the expectation of the variables for all i.

Let μ_i and σ_i be the projected mean and standard deviation (std) of class i samples on feature direction $\mathbf{w}$ respectively . The margin between any two classes as projected on $\mathbf{w}$ can be defined as

$$\mathcal{M}_{i,j} = |\mu_i - \mu_j| - b(\sigma_i + \sigma_j), \quad (3)$$

where b is a constant that quantifies the within-class spread on each side of the class center. For example, we can choose $b = 3$, as statistically very few data will fall outside 3-standard deviations. Equation (3) concerns the actual data distribution, and gives a good measure of class margin. However, it is non-linear and leads to analytically intractable formulation. To facilitate feature derivation, we need to represent the margin using the squared terms of (3) as follows:

$$\hat{\mathcal{M}}_{i,j} = |\mu_i - \mu_j|^2 - \beta(\sigma_i^2 + \sigma_j^2), \quad (\beta = b^2). \quad (4)$$

It can be shown that $\hat{\mathcal{M}}_{i,j}$ is grossly a monotonic increasing function of $\mathcal{M}_{i,j}$, since the terms under square are all non-negative. Therefore, we can replace $\mathcal{M}_{i,j}$ by $\hat{\mathcal{M}}_{i,j}$ for the margin estimation.

As suggested by name, the optimization criterion of MMDA is the average margin between the classes, i.e.

$$\begin{aligned} J &= E_{i,j}\{\hat{\mathcal{M}}_{i,j}\} \\ &= E_{i,j}\{|\mu_i - \mu_j|^2\} - \beta E_{i,j}\{\sigma_i^2 + \sigma_j^2\}. \end{aligned} \quad (5)$$

The first part of (5) can be rewritten as

$$\begin{aligned} E_{i,j}\{|\mu_i - \mu_j|^2\} &= \mathbf{w}^T E_{i,j}\{(\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T\}\mathbf{w} \\ &= \mathbf{w}^T E_{i,j}\{[(\mathbf{m}_i - \mathbf{m}) - (\mathbf{m}_j - \mathbf{m})] \cdot \\ &\qquad [(\mathbf{m}_i - \mathbf{m}) - (\mathbf{m}_j - \mathbf{m})]^T\}\mathbf{w} \end{aligned}$$

Expanding the terms, we can have

$$\begin{aligned} E_{i,j}\{|\mu_i - \mu_j|^2\} &= 2\mathbf{w}^T E_i\{(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T\}\mathbf{w} \\ &= 2\mathbf{w}^T S_B \mathbf{w}. \end{aligned} \quad (6)$$

The other two terms vanish to zero. Similarly, the second part of (5) can be reduced to

$$\begin{aligned} & \beta E_{i,j}\{\sigma_i^2 + \sigma_j^2\} = 2\beta E_i\{\sigma_i^2\} \\ = \; & 2\beta E_i\{E_k\{\mathbf{w}^T(\mathbf{x}_k - \mathbf{m}_i)(\mathbf{x}_k - \mathbf{m}_i)^T\mathbf{w}\}\} \\ = \; & 2\beta \mathbf{w}^T S_W \mathbf{w}. \quad (\mathbf{x}_k \in \mathcal{C}_i) \end{aligned} \quad (7)$$

Thus, (5) is equivalent to

$$J(\mathbf{w}) = \mathbf{w}^T(S_B - \beta S_W)\mathbf{w}, \quad (8)$$

and we call $S = S_B - \beta S_W$ the Margin Distribution Matrix.

The MMDA Features $\{\mathbf{w}_i\}$ can then be found as the unit vectors which satisfy

$$\mathbf{w} = \arg\max_{\mathbf{w}} J(\mathbf{w}). \quad (9)$$

It can be shown that $\{\mathbf{w}_i\}$ can be solved as the eigenvectors of S, and $J(\mathbf{w}_i)$ equals to the corresponding eigenvalue λ_i. S is symmetric but may not be positive. Hence, $\{\mathbf{w}_i\}$ is orthogonal and λ_i might be negative. Recall that $J(\mathbf{w}_i)$ measures the average margin between the classes on feature direction $\mathbf{w}_i$. Positive λ_i reflects the separation between the classes, while negative λ_i indicates the overlap extent between the classes. Since the spread of class data is usually larger than 1 std radius, $\beta \gg 1$ is preferred for proper estimation of margin. As a rule of thumb, one can approximately set $\beta = 9$ for sparse sample problems, assuming a data spread of 3 std on each side.

3. THE COMPUTATIONAL TRICK

MMDA does not involve inverse of S_W, and hence, can successfully deal with the sparse sample problem. But since it directly operates on the input space, we need to do eigen-decomposition on a $d \times d$ matrix. In the case of face recognition, we usually have $d \gg N$ as d is the image pixel size. For example, a 64×64 image would lead to $d = 4096$ for gray-level images. If we deal with color images, d could be even larger. Many computing machines can not easily handle eigen-decomposition on such a scale.

One way to handle this is of course to reduce the input space dimension. One can either size down the image or project it to a lower dimensional subspace. But in doing so, we may lose class-discriminant information. Another way is to take the fact that $S = S_B - \beta S_W$ is symmetric, and $rank(S) \ll d$. We can find a matrix Z such that $S = ZZ^T$, and compute the eigenvectors of S from a much smaller matrix $Z^T Z$. This is similar to how we find high-dimensional PCA features [2]. But here, complex numbers are involved when we develop the computational trick.

Let $\Phi = \{\phi_1, \phi_2, ..., \phi_C\}$ and $\Psi = \{\psi_1, \psi_2, ..., \psi_N\}$, where $\phi_i = \sqrt{\frac{N_i}{N}}(\mathbf{m}_i - \mathbf{m})$ and $\psi_k = \sqrt{\frac{1}{N}}(\mathbf{x}_k - \mathbf{m}_i)$ ($\mathbf{x}_k \in \mathcal{C}_i$). It can be verified that $S_B = \Phi\Phi^T$ and $S_W = \Psi\Psi^T$.

Hence, the margin distribution matrix S can be rewritten as $S = XX^T - YY^T$, where $X = \Phi$ and $Y = \sqrt{\beta}\Psi$. Let $Z = [X|\mathrm{i}Y]$, we have $S = ZZ^t$. Note that 't' denotes the non-conjugate transpose, for example, $(\mathrm{i}Y)^t = \mathrm{i}Y^T$ if Y is real.

Z is now a $d \times (C + N)$ complex matrix. But as S is real and symmetric, the eigenvectors and eigenvalues of S, $\{\mathbf{w}_i\}$ and $\{\lambda_i\}$, are all real. Here, we show how to compute $\{\mathbf{w}_i\}$ and $\{\lambda_i\}$ from a complex matrix Z^tZ, whose size is only $(C+N)\times(C+N)$. First, we do eigen-decomposition on Z^tZ:

$$Z^tZ\hat{\mathbf{w}}_i = \hat{\lambda}_i\hat{\mathbf{w}}_i. \tag{10}$$

This is of lower computation because, under our assumption, Z^tZ has a much smaller size than ZZ^t. Multiplying Z to both sides of (10), we have

$$ZZ^t(Z\hat{\mathbf{w}}_i) = \hat{\lambda}_i(Z\hat{\mathbf{w}}_i). \tag{11}$$

Thus, we can see that $\hat{\lambda}_i = \lambda_i$, and $Z\hat{\mathbf{w}}_i = \alpha\mathbf{w}_i$, where α might be a real or complex constant.

Because Z and Z^tZ are complex, $\hat{\mathbf{w}}_i$ and hence $Z\hat{\mathbf{w}}_i$ could be complex. Without loss of generality, we assume $Z\hat{\mathbf{w}}_i = \mathbf{a} + \mathrm{i}\mathbf{b}$, where $\mathbf{a}$ and $\mathbf{b}$ are real vectors. (11) now becomes $ZZ^t\mathbf{a}+\mathrm{i}ZZ^t\mathbf{b} = \hat{\lambda}_i\mathbf{a}+\mathrm{i}\hat{\lambda}_i\mathbf{b}$, which suggests that

$$ZZ^t\mathbf{a} = \hat{\lambda}_i\mathbf{a} \quad \text{and} \quad ZZ^t\mathbf{b} = \hat{\lambda}_i\mathbf{b}. \tag{12}$$

Therefore, $\mathbf{a}$ and $\mathbf{b}$ can either be a zero vector, or a real multiple of the eigenvector $\mathbf{w}_i$. Either case, we can compute $\mathbf{w}_i$ from $Z\hat{\mathbf{w}}_i$ as

$$\mathbf{w}_i = \mathbf{v}/norm(\mathbf{v}) \quad (\mathbf{v} = real(Z\hat{\mathbf{w}}_i) + imag(Z\hat{\mathbf{w}}_i).) \tag{13}$$

Equations (10) and (13) together give us an efficient way to extract the MMDA features directly from a high-dimensional input space. This will importantly preserve the features with highest discrimination/separation. Whereas, use of PCA for dimension reduction as in many LDA implementations is generally not optimal in this aspect.

4. EXPERIMENTAL RESULTS

To evaluate the performance of MMDA for face recognition problems, we conduct intensive tests on both the ORL [9] and the UMIST [10] face databases. The ORL database consists of 40 persons, each with 10 images. The UMIST database has 575 images of 20 persons, and is of more drastic pose variation as compared to ORL. All the face images in use are 112×92 and we use the proposed computational trick to extract the MMDA features. To make the results less dependant of the training samples, all the results reported are averaged over 100 random runs. Here, we focus on the first $C - 1$ features of MMDA, i.e. 39 for ORL and 19 for UMIST, since this is the size limit of the LDA features that we compare to. Classification is done by the simple nearest neighbor method with the L_2 norm-based distance.

4.1. The overall performance

First, we evaluate the performance of MMDA under different β and compare it to that of LDA. Fig. 1 summarizes the results of the two methods for ORL and UMIST databases at the optimal feature size (i.e. the number of features that achieves the highest average accuracy). The upper curve shows the accuracy rate with the performance deviation of 1 std radius indicated, while the lower bar graph gives the corresponding optimal feature size. For MMDA plots, we highlight the three points where $\beta = -1$ (PCA equivalent), $\beta = 1$ (MMC) and $\beta = 9$ (Our proposed setting). For LDA plots, the K_{pca} setting that achieves the highest accuracy is highlighted.

It can be seen that the performance of MMDA appears stabled when $\beta \gg 1$, while the performance of LDA varies much with K_{pca}. This confirms with our earlier claim that MMDA is very stable for sparse sample problems. In terms of accuracy, MMDA is also very competitive with LDA. Table 1 lists the accuracy rates of the two methods at the highlighted points. We can see that at a large β, e.g. $\beta = 9$, MMDA achieves better performance even than the best rates of LDA. In practice, it is hard to estimate the optimal K_{pca} setting based only on the training images. Hence, MMDA is more preferred as its performance does not vary much when $\beta \geq 9$. Another observation is that LDA can sometimes achieve similar performance as MMDA but with fewer number of features. For example, in the ORL tests, the best feature size is 23 for $K_{pca} = 40$, and 38 for $\beta = 9$. This is because the MMDA features are constrained to be orthogonal to each other, while the LDA features are not. Hence, the LDA features can be more effective for classification for some kind of distributions. But MMDA can compensate this with a few more features added. Considering its performance gain and stableness, this is a justifiable cost.

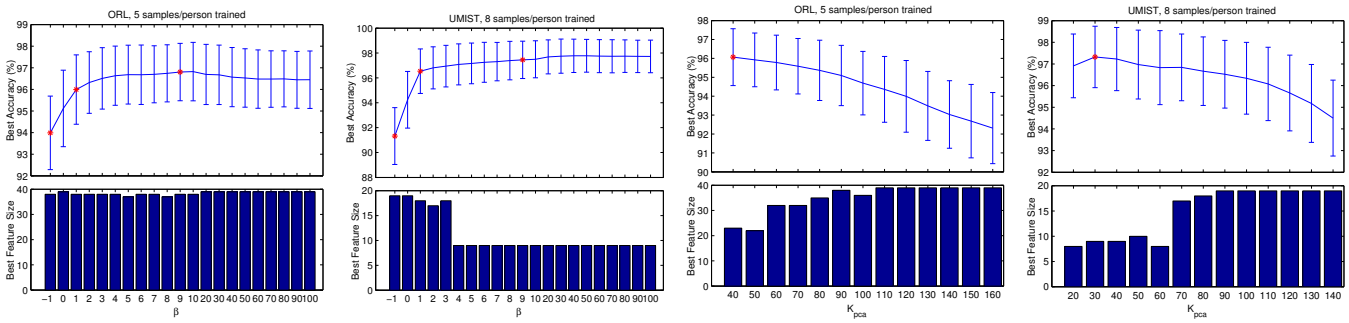
Table 1. Classification accuracy (%) of MMDA and LDA at the highlighted points.

	MMDA			LDA
	$\beta = -1$	$\beta = 1$	$\beta = 9$	K^*_{pca}
ORL	93.99/1.70	96.00/1.61	**96.81**/1.33	96.07/1.50
UMIST	91.33/2.28	96.54/1.79	**97.46**/1.50	97.33/1.42

The value after / is the performance deviation (std) over 100 runs.

4.2. The β setting effects

Looking into the β setting itself, we can also see that the performance is enhanced greatly from the PCA equivalent to MMC, and then to where $\beta = 9$. After that point, the change is very small. For ORL, the performance slightly drops when $\beta > 9$, because a larger β may over-punish the feature space searching, and hence miss some small margin features. But for UMIST, where the class distribution is known to be more sparse, a $\beta > 9$ actually slightly boosts



(a) (b) (c) (d)

Fig. 1. Recognition performance of (a)(b) MMDA, and (c)(d) LDA for different β or K_{pca}. (a)(c): ORL, 5 samples/person trained. (b)(d): UMIST, 8 samples/person trained.

up the performance. Also, $\beta = 1$ and $\beta = 9$ may generate very different sets of features, though their performance difference is not very large. This is evident by the UMIST test where the best feature size is 18 for $\beta = 1$ and 9 for $\beta = 9$. Obviously, in this situation, the latter setting provides better regularization, and produce more class-discriminant features.

4.3. Varying the feature size

To have a closer look of how MMDA and LDA performances vary with the feature size, we show a typical plot in Fig. 2 using the ORL database. An interesting point in Fig. 2 is that the first few features of MMDA can actually be more useful than those of the LDA ones. Hence, comparing to the fisher's discriminant criteria, margin is perhaps a better measure of the feature usefulness for classification purpose, as it more closely reflects the extent of non-overlap among the classes.

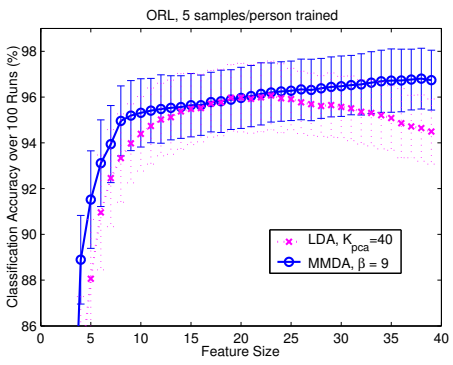


Fig. 2. Performance of MMDA and LDA at varied feature size.

5. CONCLUSION

In this paper, we propose a new type of discriminant analysis called MMDA. It extracts features by maximizing the average margin between the classes, and has a class-spread parameter β which counts for sample distribution. The MMDA features can be directly obtained from the input space, and a computational trick is proposed to handle high dimensional data. We compare MMDA to LDA for face recognition problems with ORL and UMIST databases. The results show that MMDA is a good replacement of LDA for sparse sample problems.

6. REFERENCES

[1] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, pp. 711-720, 1997.

[2] H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional data with application to face recognition", *Pattern Recognition*, vol. 34, pp. 2067-2070, 2001.

[3] M.-H. Yang, "Kernel Eigenfaces vs. Kernel Fisherfaces: Face Recognition Using Kernel Methods", *Proc. of the Fifth Inter. Conf. on Automatic Face and Gesture Recognition (FG 2002)*, pp. 215-220, 2002.

[4] J. Lu, K.N. Plataniotis, and A.N. Venetsanopoulos, "Regularized Discriminant Analysis For the Small Sample Size Problem in Face Recognition", *Pattern Recognition Letter*, vol. 24(16), pp. 3079-3087, 2003.

[5] W. Tian, M. Barbero, Z. Gu, S. Lee, "Image classification by the foley-sammon transform", *Opt. Eng.*, vol. 25(7), pp. 834-840, 1986.

[6] Z.-Q. Hong and J.-Y. Yang, "Optimal discriminant plane for a small number of samples and design method of classifier on the plane", *Pattern recognition*, vol. 24(4), pp. 317-324, 1991.

[7] J.H. Friedman, "Regularized discriminant analysis", *J. of the American Statistical Assoc.*, vol. 84, pp. 165-175, 1989.

[8] H. Li, T. Jiang, and K. Zhang, "Efficient and Robust Feature Extraction by Maximum Margin Criterion", *In Advances in Neural Information Processing Systems 16*, 2003.

[9] The ORL face database, [online], available: http://www.uk.research.att.com/facedatabase.html

[10] The UMIST face database, [online], available: http://images.ee.umist.ac.uk/danny/database.html