Analysis On Fisher Discriminant Criterion And Linear Separability Of Feature Space

Yong Xu
Department of Computer Science & Technology, Shenzhen Graduate School, Harbin Institute of Technology
Shenzhen, China 518055
laterfall@hitsz.edu.cn

Guangming Lu
The Graduate School at Shenzhen, Tsinghua University
Shenzhen, China 518055
luguangm@gmail.com

Abstract

For feature extraction resulted from Fisher discriminant analysis (FDA), it is expected that the optimal feature space is as low-dimensional as possible while its linear separability among different classes is as large as possible. Note that the existing theoretical expectation on the optimal feature dimensionality may contradict with experimental results. Due to this, we address the optimal feature dimensionality problem with this paper. The multi-dimension Fisher criterion is used to measure the linear separability of the feature space obtained using FDA and to analyze the optimal feature dimensionality problem. We also attempt to answer the question "what kind of real-world application is FDA competent for". Theoretical analysis shows that the genuine optimal feature dimensionality should be lower than that presented by Jin et al. A number of experiments illustrate that the proposed optimal feature extraction does have advantages.

Key words: Multi-dimension Fisher criterion; Feature extraction; Linear separability

1. Introduction

The basic target of the widely used Fisher discriminant analysis (FDA) [1-15] is to seek a transforming axis which is able to transform samples into ones with maximal linear separability. The transforming axis that leads to the best linear separability is called optimal transforming axis. Generally, besides the first optimal transforming axis is available; suboptimal transforming axes can be also obtained. In practice, an N dimensional sample space has N available transforming axes in total. Note that extended Fisher discriminant analysis (KFDA) [16-20], which is derived from FDA methodology, has also received much attention.

A number of studies show that if samples are transformed into a lower-dimensional subspace, higher classification accuracy may be expected [6]. For real-world applications, if FDA is required to extract low-dimensional features of samples, the following problems should be addressed. The first problem is which transforming axes should be used for feature extraction. The second problem, which is also called optimal feature dimensionality problem, is what is the optimal feature dimensionality.

Jin et al. [6] claimed that all the FDA discriminant vectors associated with positive Fisher criterion values are helpful to obtain useful classification information in extracting features of samples. As a result, they insisted that all these discriminant vectors be used for feature extraction. Generally, there are L-1 available positive Fisher criterion values, thus samples can be transformed into a L-1 dimensional space using the transforming axes associated with these criterion values. This is the first attempt to relate the optimal number of transforming axes to the number of sample classes. On the other hand, though these L-1 transforming axes associated with the positive criterion values were expected to achieve the best classification performance, a number of experiments did not meet this expectation. For example, refs. [7,8,9] all indicated that the feature space associated with the highest accuracy did not has the dimensionality of L-1. In some cases, the classification accuracy of a feature space with the dimensionality smaller than L-1 may be higher than that of the L-1 dimensional feature space [10]. In addition, the following different statement is also available: for the small training

1-4244-0605-6/06/$20.00 ©2006 IEEE.

sample set, the optimal feature dimensionality should be greater than L-1 [11]. It appears that up to now there is no satisfactory solution to the optimal feature dimensionality problem.

The remainder of this paper is organized as follows. In section 2 we propose the notion of the multi-dimension Fisher criterion. We also analyze the theoretical properties of FDA. In section 3, we analyze the physical meaning of the multi-dimension Fisher criterion. In section 4, we conduct several experiments to illustrate our theoretical analysis. Section 5 offers our conclusion.

2. The multi-dimension Fisher criterion

The problem to seek the optimal transforming axis of FDA is identical to the one that maximizes the following Fisher criterion [3]: $f(w)=\frac{w^T S_b w}{w^T S_w w}$, where S_b, S_w are the between-class scatter matrix and the within-class scatter matrix, respectively. S_b, S_w are respectively defined as follows:

$$S_b=\sum_{i=1}^{L} p(\omega_i)(m_i-m_0)(m_i-m_0)^T, \quad (1)$$

$$S_w=\sum_{i=1}^{L} p(\omega_i)E[(y-m_i)(y-m_i)^T \mid \omega_i], \quad (2)$$

where $\omega_1,\omega_2,...\omega_L$ denote L classes, m_i the expectation of ω_i, $p(\omega_i)$ the prior of ω_i, and m_0 the expectation of the total samples. Suppose that the dimensionality of original samples is N. As a result, S_b and S_w are both $N\times N$ matrices. In practice, the problem to determine multi transforming axes is equivalent to the one which aims to maximize the following Fisher criterion [12]:

$$F(W)=\frac{|W^T S_b W|}{|W^T S_w W|}. \quad (3)$$

Since multi transforming axes are required, the criterion (3) can be named multi-dimension Fisher criterion. The columns of the matrix W are composed of the transforming axes. If the W maximizes the criterion (3), it can be called optimal transforming matrix. When there are d transforming axes, the W is an $N\times d$ matrix. Consequently, a sample y can be reformed to be a d dimensional vector by the transform $z=W^T y$.

It is known that column vectors of the W should consist of eigenvectors of the following general eigenequation

$$S_b x=\lambda S_w x. \quad (4)$$

Suppose that eigen-values of (4) are $\lambda_1\ge\lambda_2\ge,...,\ge\lambda_N\ge 0$. Suppose that S_w is a positive definite matrix. Then Eq. (4) can be solved directly We also assume that only distance classifiers are used for classifying samples.

The number d of the column vectors of W may vary from 1 to N. In this paper the feature extraction associated with the best linear separability is called optimal feature extraction. In practice, two quantitative indicators are available for the FDA transform. The first quantitative indicator is the Fisher criterion value of a transforming axis. The greater the value is, the more linearly separable the features of different classes obtained using the transforming axis is. The second quantitative indicator is the number of the transforming axes which can lead to the best linear separability. We present our analysis on the second problem via the following theorems. Hereafter we define that the transforming matrix W is composed of the eigenvectors associated with the first c largest eigen-values of (4).

Theorem 1. $\lambda_1\lambda_2...\lambda_c$ is the value of the multi-dimension Fisher criterion associated with W.

Proof.

Let $x_1,x_2...x_c$ be the eigenvectors associated with the first c largest eigen-values of Eq.(4), respectively. As a consequence, we have

$$S_b x_i=\lambda_i S_w x_i, 1\le i\le c. \quad (5)$$

It follows by $W=[x_1 \quad x_2 \quad . \quad . \quad . \quad x_c]$ that

$$(W^T S_b W)_{ij}=x_i^T S_b x_j,$$
$$(W^T S_w W)_{ij}=x_i^T S_w x_j. \quad (6)$$

Note that if $\lambda_i\ne\lambda_j$ which is almost always satisfied for real-world applications, the following formulas must be certain: $x_i^T S_b x_j=0, x_i^T S_w x_j=0, i\ne j, 1\le i,j\le N$. As a result, we can be led to the following equalities:

$$|W^T S_b W|=(x_1^T S_b x_1)(x_2^T S_b x_2)...(x_c^T S_b x_c), \quad (7)$$

$|W^T S_w W| = (x_1^T S_w x_1)(x_2^T S_w x_2)...(x_c^T S_w x_c)$. (8)

On the other hand, the following formula is certain:

$$x_i^T S_b x_i = \lambda_i x_i^T S_w x_i, 1 \le i \le N \quad . \quad (9)$$

Substituting (9) into (7) yields $\frac{|W^T S_b W|}{|W^T S_w W|} = \lambda_1 \lambda_2 ... \lambda_c$, which says that the value of the multi-dimension Fisher criterion is as follows:

$$F(W) = \lambda_1 \lambda_2 ... \lambda_c . \quad (10)$$

Moreover, we have the following theorems.

Theorem 2. Under the condition $\lambda_1, \lambda_2, ..., \lambda_c \ge 1, \lambda_{c+1}, \lambda_{c+2}, ..., \lambda_N < 1$, the multi-dimension Fisher criterion associated with the transforming matrix W reaches its maximum value.

Proof.

We will use the reduction to absurdity to demonstrate this theorem. Firstly, suppose that $\lambda_1, \lambda_2, ..., \lambda_c \ge 1$ is not satisfied and instead there are $\lambda_c < 1$, $\lambda_1, \lambda_2, ..., \lambda_{c-1} \ge 1$ and $\lambda_{c+1}, \lambda_{c+2}, ..., \lambda_N < 1$. As a consequence, it is certain that $\lambda_1 \lambda_2 ... \lambda_c < \lambda_1 \lambda_2 ... \lambda_{c-1}$. In other words, based on the above supposition, the W will not result in the maximum multi-dimension Fisher criterion value.

Secondly, assume that $\lambda_{c+1}, \lambda_{c+2}, ..., \lambda_N < 1$ is not satisfied and there are $\lambda_{c+1} \ge 1$ and $\lambda_{c+2}, ..., \lambda_N < 1$, $\lambda_1, \lambda_2, ..., \lambda_c \ge 1$. As a result, we have $\lambda_1 \lambda_2 ... \lambda_c \le \lambda_1 \lambda_2 ... \lambda_c \lambda_{c+1}$, which also shows that the multi-dimension Fisher criterion associated with the W does not reach its maximum value.

The demonstration above indicates that under the condition of $\lambda_1, \lambda_2, ..., \lambda_c \ge 1, \lambda_{c+1}, \lambda_{c+2}, ..., \lambda_N < 1$ the multi-dimension Fisher criterion associated with W must arrive at the maximum value.

In fact, under the condition of theorem 2, $W = [x_1 \quad x_2 \quad . \quad . \quad . \quad x_c]$ is the optimal transforming matrix, for it will result in the maximal multi-dimension Fisher criterion value, which implies that the corresponding feature space has greater linear separability than other feature spaces. Hence, c is the optimal feature dimensionality. On the other hand, we can know that the optimal feature extraction specified by Jin does not correspond to the maximum value of the multi-dimension Fisher criterion.

Theorem 3. The optimal feature dimensionality proposed by Jin is generally greater than that specified in this paper.

Proof.

It is clear that nonzero eigen-values are more than those greater than 1. Thus, we have L-1 $\le c$. Thus, we may say that the optimal feature dimensionality proposed by Jin is generally lower than that specified by us.

3. More discussion on the optimal feature dimensionality

3.1 Threshold of Fisher criterion and its physical meaning

In practice, to take 1 as the threshold of Fisher criterion has the following physical meaning: if Fisher criterion value is greater than 1, the average distance between sample features from different classes will be greater than that between sample features in the same class. As a result, samples are suited to be classified by a distance classifier and acceptable classification performance can be expected. Now we illustrate this by taking the two-class problem as an example.

Suppose that the eigenvector associate with the eigen-value λ of $S_b x = \lambda S_w x$ is used for feature extraction, then the one-dimensional feature obtained by the feature extraction process is $z = y^T x$, where y means an arbitrary sample. Let z_1, z_2 be the mean-vectors of the two classes, respectively. For the feature space, the average squared distance between the means of the two classes can be formulated as follows: $(z_1 - z_2)^2 = x^T S_b x$. The greater the value of the formula is, the more severely the features of different categories vary and the greater the distance between the means of the two classes is.

Similarly, the average of the variance of the sample features in the same class is $\frac{1}{2}\frac{1}{n}\sum_{i=1}^{2}\sum_{j \in \omega_i}(z_{ji} - z_i)^2 = x^T S_w x$, where z_{ji} is the feature of the j th sample of the i th category, and n the number of the samples in each class. The smaller this average variance is, the smaller the difference between sample features in the same class is. Clearly, large $x^T S_b x / x^T S_w x = \lambda$ means great linear separability between sample features from different categories.

Therefore, the Fisher criterion value smaller than 1 indicates that the distance between sample features from the same category is statistically larger than that between sample features from different classes. Consequently, the corresponding transforming axis is not of significance for feature extraction and consequent classification. On the other hand, a transforming axis associated with Fisher criterion value greater than 1 is favorable.

3.2 Analysis on multi-dimensional Fisher criterion

The scheme of using all the eigenvectors associated with nonzero eigen-values as transforming axes is subject to the adding principle of single Fisher criterion value. That is, it is considered that the larger the sum of all the positive single Fisher criterion values, the greater the linear separability of a feature space is.

The multi-dimension criterion proposed in this paper measures the linear separability of the feature space using the multiplication principle. This principle is consistent with the following fact. A Fisher criterion value less than 1 means that the between-class distance is smaller than the within-class distance. In practice, the value of the multi-dimension Fisher criterion descends while this kind of transforming axis is also used as transforming axis. Thus, the multiplication principle directly relate the multi-dimension Fisher criterion value to the linear separability measure of the feature space.

The multi-dimension Fisher criterion can measure and compare the linear separability of different dimensional feature spaces with the following conclusions.

(1) If all eigen-values of the eigen-equation are larger than 1, the maximums value of the multi-dimension Fisher criterion generally coincides with the transforming matrix consisting of all available eigenvectors. Hence, among all the feature spaces the one associated with this transforming matrix has the maximum linear separability. In this case, it seems that feature extraction performed using FDA is unuseful to improve classification performance.

(2) For the real-world case that samples are very high-dimensional and the number of sample categories is small, the between-scatter matrix will have a number of zero eigen-values which are usually computationally nonzero. As a consequence, a lower-dimensional subspace obtained using FDA-based transform may have a large multi-dimension Fisher criterion value. Also, high classification accuracy may be available.

4 · Experiments

4.1 Experiment on face images

We transformed the images of AR face database into grey-level ones and cropped them to obtain images each having 60×60 pixels. We performed the first face recognition experiment using parts of typical images per individual as training samples. That is, the first, fifth, eighth, eleventh and fourteenth images of each individual, which characterize all typical variations of an individual face (as shown in the first row of Fig. 1) and are called typical images, were used as training samples, while the others were employed as test samples. The classification process is called classification on typical facial variations. Table 1 tells us that our approach and Jin's approach proposed in ref.[6] obtain accuracies of 75.6% and 75.4%, respectively.

Face classification under the condition of varying facial expression was also performed, followed by the face recognition using face images with varying lighting condition. For each of these two cases, two samples of every class were used as members of the training sample set, while six samples from the same category would be taken as test samples. Both the cases selected the first and fourteenth images of each class as training samples. In face classification task with varying facial expression, the test sample set included the second, third, fourth, fifteenth, sixteenth and seventeenth samples of every class, which contains expressions such as smile, anger and scream. For face recognition using images with varying lighting, the test sample set consisted of the fifth, sixth, seventh, eighteenth, nineteenth and twentieth images of every class.

From Table 1, we can see that our approach and the approach in ref.[6] obtained two comparable accuracies, 85.3% and 85.4%, when classifying faces with varying expression. For face recognition under the condition of varying lighting, accuracies of our approach and the approach in ref. [6] were 82.9% and 82.7%, respectively. Moreover, our approach used fewer transforming axes than the approach in ref. [6] to extract features of samples. Therefore, it is certain that our approach took less time than the approach of ref.[6] to perform feature extraction and consequent classification.

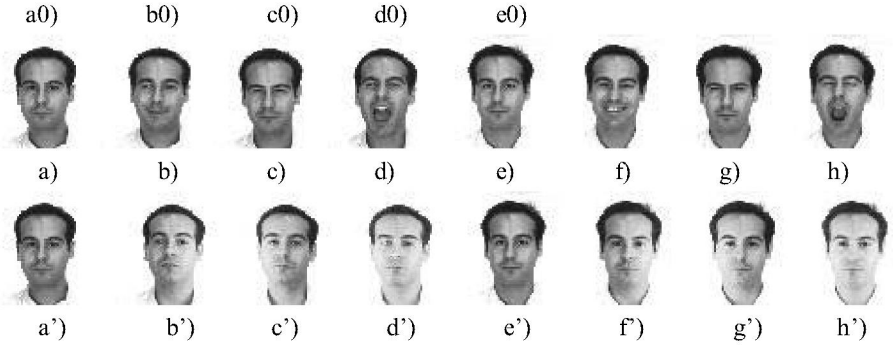

Fig 1. The first row shows some samples of an individual in AR database. These images characterize all typical facial variations of an individual. The second row shows images used for face classification using images with varying expression. a) and e) were selected as training samples, while b),c),d), f) ,g) and h) were included in the test sample set. The third row shows images used for face recognition on images with varying lighting. a') and e') were selected as training samples, while b'),c'),d'), f') ,g') and h') were taken as test samples. Note that a) and a') denote the same image, while e) and e') mean another identical image.

Table 1. Experimental result on AR face images

	Our approach	The approach in ref.[6]
Classification accuracy of typical facial variations	75.6%	75.4%
Accuracy under the condition of varying expression	85.3%	85.4%
Accuracy under the condition of varying lighting	82.9%	82.7%

4.2 Experiment on palmprint images

We collected 300 left palmprint images from 50 subjects, each having 6 palmprint images. The training set consists of the first three images of every subject, while the test sample set is composed of the other images. Table 2 shows that both our approach and the approach of ref.[6] produced the same accuracy, 83.3%. In addition, because more transforming axes were used by the feature extraction procedure proposed in ref.[6], the approach in ref.[6] would take longer time than our approach to extract features of samples and to classify them.

Table 2. Classification performance on palmprint image database

	Our approach	Approach in ref.[6]
Classification accuracy	83.3%	83.3%

5. Conclusion

This paper clearly presents the physical meaning of the single Fisher criterion and multi-dimension Fisher criterion. While the single Fisher criterion identifies the discriminant performance of a transforming axis, the multi-dimension Fisher criterion measure the linear separability of the feature space obtained using multi transforming axes. The multi-dimension Fisher criterion may be also used to compare the linear separability of two different-dimensional feature spaces. Optimal transforming matrix, defined as the one that is associated with the maximum multi-dimension Fisher criterion, is able to result in the maximal linear separability. It appears the maximum multi-dimension Fisher value may coincide with the best classification performance. Moreover, according to the multi-dimension Fisher criterion, it is probably that FDA is more suitable for feature extraction in the case with high-dimensional samples and a few sample classes than opposite cases.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No. 60472060, No. 60473039, and No. 60620160097), Natural Science Foundation of Guangdong province (No. 06300862).

References

[1] R.A. Fisher, "The use of multiple measures in taxonomic problems", Ann. Eugenics, 7:

(1936) 179-188.
[2] Y. Xu, J.-Y. Yang, and Z. Jin, "Theory analysis on FSLDA and ULDA". Pattern Recognition, 36(12) (2003) 3031-3033.
[3] Y. Xu, J.-Y. Yang, and Z. Jin, "A novel method for Fisher discriminant Analysis". Pattern Recognition, 37 (2) (2004) 381-384.
[4] Z. Jin, J.-Y. Yang, Z. Lou, Z. Tang, and Z. Hu, "A theorem on the uncorrelated optimal discrimination vectors", Pattern Recognition., 34 (10) (2001) 2041–2047.
[5] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenface vs. Fisherface: recognition using class specific linear projection", IEEE Trans. Pattern Anal. And Mach. Intelligence, 19(10) (1997) 711-720.
[6] Z. Jin, J.-Y. Yang, Z. Hu, and Z. Lou, "Face recognition based on the uncorrelated discrimination transformation", Pattern Recognition, 34(7) (2001) 1405–1416.
[7] X.-Y. Jing , H.-S. Wong, D. Zhang, and Y.-Y. Tang, "An uncorrelated fisherface approach", Neurocomputing, Vol. 67 (2005) 328-334.
[8] X.-Y. Jing, D. Zhang, and Y.-Y. Tang, "An Improved LDA Approach", IEEE Trans.Systems, Man and Cybernetics, Part B, 34(5) (2004) 1942 −1951.
[9] F.-B. Chen, S.-L. Zhang, X.-M. Gao, and, J.-Y. Yang, "Theory of Fisher linear discriminant analysis for small sample size problem and its verification", No. 8: (2005) 984~991.
[10] J. Yang, J.-Y. Yang, and D. Zhang, "What's wrong with Fisher criterion?". Pattern Recognition, 35(12) (2002) 2665-2668
[11] J. Yang, and J.-Y. Yang, "Why can LDA be performed in PCA transformed space?". Pattern Recognition, 36 (2) (2003) 563-566.
[12] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2nd ed. New York: Wiley, 2001.
[13] H. Yu, and J. Yang, "A direct LDA algorithm for high-dimensional data with application to face recognition". Pattern Recognition, 34 (10) (2001) 2067-2070
[14] T. Hastie, and R. Tibshirani , "Discriminant adaptive nearest neighbor classification", IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(6) (1996) 607-616.
[15] M. Bressan, and J. Vitrià, "Nonparametric discriminant analysis and nearest neighbor classification", Pattern Recognition Letters 24 (2003) 2743 - 2749.
[16] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller, "Fisher discriminant analysis with kernel". In YHHu, J.Larsen, and S.Wilson, E.~Douglas, editors, Neural Networks for Signal Processing, pages 41--48.
[17]Y. Xu, J.-Y. Yang, and J. Yang, "A reformative kernel Fisher discriminant analysis", Pattern Recognition, 37 (6) (2004) 1299-1302.
[18] Y. Xu, J.-Y. Yang, J. Lu, and D.-J. Yu, "An efficient renovation on kernel Fisher discriminant analysis and face recognition experiments". Pattern Recognition, 37 (10) (2004) 2091-2094.
[19] J. Yang , A. F. Frangi, J.-Y Yang, D. Zhang, and Z. Jin, "KPCA plus LDA: a complete kernel Fisher discriminant framework for feature extraction and recognition". IEEE Trans. Pattern Anal. Mach. Intell. , 27(2) (2005) 230-244.
[20] Y. Xu, D .Zhang, Z. Jin, M. Li, and J.-Y. Yang, "A fast kernel-based nonlinear discriminant analysis for multi-class problems", Pattern Recognition, 39(6) (2006) 1026-1033.