Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

PERGAMON

Pattern Recognition 37 (2004) 2091–2094

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Rapid and Brief Communication

An efficient renovation on kernel Fisher discriminant analysis and face recognition experiments

Yong Xu*, Jing-yu Yang, Jianfeng Lu, Dong-jun Yu

Department of Computer Science, Nanjing University of Science and Technology, Nanjing, Jiangsu 210094, China

Received 30 January 2004; accepted 13 February 2004

Abstract

A reformative kernel algorithm, which can deal with two-class problems as well as those with more than two classes, on Fisher discriminant analysis is proposed. In the novel algorithm the supposition that in feature space discriminant vector can be approximated by some linear combination of a part of training samples, called "significant nodes", is made. If the "significant nodes" are found out, the novel algorithm on kernel Fisher discriminant analysis will be superior to the naive one in classification efficiency. In this paper, a recursive algorithm for selecting "significant nodes", is developed in detail. Experiments show that the novel algorithm is effective and much efficient in classifying.

Keywords: Fisher discriminant analysis; Kernel trick; Pattern recognition; Feature space

1. Introduction

Kernel Fisher discriminant analysis has been applied to many pattern recognition problems and its good performance is available [1,2]. Its basic idea can be described that input space is mapped into some feature space (usually nonlinear space), consequently Fisher discriminant analysis is performed in the feature space. It is notable that in kernel Fisher discriminant analysis any explicit mapping is not necessary, because kernel trick is introduced [1]. However, it is well known that the classification efficiency of naive kernel Fisher discriminant analysis descends while the number of training samples increases, and as a result its application for some practical problems with many training samples may be very time consuming and even impossible. Supposing that in feature space discriminant vector can be approximated by some expression appearing as linear combination of a part of training samples, some algorithms have been proposed to improve classification efficiency of kernel methods [2,3]. Nevertheless, these algorithms are only developed for two-class problems. In this paper, a reformative algorithm on kernel Fisher discriminant analysis is developed, which is valid for two-class problems, as well as those with more than two classes. We call the training samples, whose linear combination may approximate well the discriminant vectors in feature space, "significant nodes". A cursive algorithm is presented to select "significant nodes". Although the algorithm for selecting "significant nodes" may be not optimal, it is feasible and reasonable, and its computational cost is acceptable. This paper is organized as follows. The next section will introduce kernel Fisher discriminant analysis and its algorithm, while Section 3 will propose our reformative algorithm. Experiments results will be given in the last section.

2. Kernel Fisher discriminant analysis

Let $\{x_i\}$ denote the input space. Suppose that the feature space is F and the corresponding nonlinear function is ϕ, i.e. $\phi(x_i) \in F$. Consequently, in the feature space F Fisher criterion is defined by

$$J(w) = \frac{w' S_b^{\phi} w}{w' S_w^{\phi} w}, \tag{1}$$

* Corresponding author. Tel.: +86-25-882-7574; fax: +86-25-431-5510.
E-mail addresses: laterfall@sohu.com (Y. Xu), Yangjy@mail.njust.edu.cn (J.-y. Yang).

0031-3203/$30.00 © 2004 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2004.02.016

where w is discriminant vector, S_b^ϕ and S_w^ϕ are between-class scatter matrix and within-class scatter matrix, respectively. Suppose that there are L classes, and the number of samples in the ith class is l_i. In addition, we define that the total number of samples is l and $x_j^i = 1, 2, \ldots, l_i$ denotes the jth sample in the ith class. If the prior probabilities of the L classes are equal, then S_b^ϕ and S_w^ϕ can be expressed as

$$S_b^\phi = \sum_{i=1,\ldots,L} (m_i^\phi - m_0^\phi)(m_i^\phi - m_0^\phi)', \tag{2}$$

$$S_w^\phi = \sum_{i=1,\ldots,L} \sum_{j=1,\ldots,l_i} (\phi(x_j^i) - m_i^\phi)(\phi(x_j^i) - m_i^\phi)', \tag{3}$$

where $m_i^\phi = 1/l_i \sum_{j=1,\ldots,l_i} \phi(x_j^i), m_0 = 1/L \sum_{i=1,\ldots,L} m_i^\phi$. The theory of reproducing kernels says that the discriminant vectors of feature space are in the space spanned by all the training samples [1]. So, w can be formulated by

$$w = \sum_{i=1,\ldots,l} \alpha_i \phi(x_i). \tag{4}$$

Now we substitute kernel function $k(x_i, x_j)$ for dot production $\phi(x_i) \cdot \phi(x_j)$. If $M_i \in R^{l\times 1}$ and $N \in R^{l\times l}$ are defined as

$$(M_i)_j = (1/l_i) \sum_{k=1,\ldots,l_i} k(x_j, x_k^i),$$

$$j = 1, 2, \ldots, l,\ i = 1, 2, \ldots, L, \tag{5}$$

$$N = \sum_{i=1,\ldots,L} K_i(I - I_{l_i})K_i', \tag{6}$$

where I is the identity, I_{l_i} is a $l_i \times l_i$ matrix and each element is $1/l_i, K_i$ is a $l \times l_i$ matrix, $(K_n)_{i,j} = k(x_i, x_j^n),\ i = 1, 2, \ldots, l,\ j = 1, 2, \ldots, l_n,\ n = 1, 2, \ldots, L$, Fisher criterion in the feature space will be expressed by [1]

$$J(\alpha) = \frac{\alpha' M \alpha}{\alpha' N \alpha}, \tag{7}$$

where $\alpha = [\alpha_1 \ \cdots \ \alpha_l]',\ M = \sum_{i=1,\ldots,L} (M_i - M_0)(M_i - M_0)',\ M_0 = (1/L)\sum_{i=1,\ldots,L} M_i$. Consequently, the problem for obtaining w is transformed into one for solving eigenvectors α, corresponding to nonzero eigenvalues of the eigenequation (8).

$$M\alpha = \lambda N\alpha. \tag{8}$$

3. The reformative algorithm on kernel Fisher discriminant analysis

In feature space discriminant vector can be approximated by some linear combination of "significant nodes", consequently, very efficient algorithms for kernel Fisher discriminant analysis are developed [2,3]. However, these algorithms are only for two-class problems. In this section we focus on multi-class problems and developing the corresponding algorithm, and the key is to develop the procedure to select "significant nodes".

According to Fisher's idea, the larger some nonzero eigenvalue of Eq. (8) is, the better the corresponding eigenvector is to be taken as discriminant vector. Generally, for multi-class problems with more than two classes, Eq. (8) corresponds to more than one nonzero eigenvalues. So, we select "significant nodes" according to the summation of nonzero eigenvalues. In other words, for different training samples we compute corresponding summation of nonzero eigenvalues of Eq. (8), and take the training samples, corresponding to the maximum summation, as "significant nodes". Based on the rule, the following procedure is proposed.

3.1. Algorithm for selecting "significant nodes"

3.1.1. Selecting the first "significant node"

For each training sample $x_i,\ i = 1, 2, \ldots, l$, corresponding M_i, K_i, M and N are computed. Obviously, here M_i, M and N are all scalars, for example $M_j = 1/l_j \sum_{k=1,\ldots,l_j} k(x_i, x_k^j)$. Consequently $\lambda_i = M/N$ is computed. After the above computation is accomplished for all the training samples, the sample corresponding to the maximum λ_i is taken as the first "significant node", denoted by x_1^o.

3.1.2. Selecting the sth "significant node"

Suppose $s-1$ samples have been selected as "significant nodes", denoted by $x_1^o, x_2^o, \ldots, x_{s-1}^o$, then selecting for the sth "significant node" will be carried out according to the following algorithm. Each sample $x,\ x \in \{x_i,\ i = 1, 2, \ldots, l\}$ and $x \notin \{x_j^0, j = 1, 2, \ldots, s-1\}$ will be considered in the procedure. When a new sample x is being considered, corresponding M_i, K_i can be formulated as follows:

$$M_i = \begin{bmatrix} M_i^0 \\ a_i \end{bmatrix}, \quad K_i = \begin{bmatrix} K_i^0 \\ k_{new}^i \end{bmatrix}, \quad i = 1, 2, \ldots, L, \tag{9}$$

where $a_i = 1/l_i \sum_{j=1,\ldots,l_i} k(x, x_j^i), k_{new}^i = [k(x, x_1^i), k(x, x_2^i) \cdots k(x, x_{l_i}^i)],\ M_i^0, K_i^0$ are M_i and K_i corresponding to the previous $s-1$ significant nodes", respectively. If the inverse matrix of N exists, Eq. (8) will be identical to eigenequation $N^{-1}M\alpha = \lambda\alpha$. To avoid the problem that solving the inverse matrix of singular matrix, eigenequation $(N + \mu I)^{-1}M\alpha = \lambda\alpha$ is often adopted instead, where μ is a positive constant. We define N_1, as

$$N_1 = \sum_{i=1,\ldots,L} K_i(I - I_{l_i})K_i' + \mu I \tag{10}$$

then we rewrite N_1 as

$$N_1 = \begin{bmatrix} N_1^0 & u \\ u' & \gamma \end{bmatrix}, \tag{11}$$

where $\gamma = \sum_{i=1,\ldots,L} k_{new}^i (I - I_{l_i})(k_{new}^i)' + \mu,\ u = \sum_{i=1,\ldots,L} K_i^0 (I - I_{l_i})(k_{new}^i)'$, N_1^0 is the N_1 corresponding to the previous $s-1$ "significant nodes". N_1 is symmetric, so N_1^{-1} can be

obtained by the formulation [2]

$$N_1^{-1} = \begin{bmatrix} (N_1^0)^{-1} + \frac{1}{\rho} zz' & -\frac{1}{\rho} z \\ -\frac{1}{\rho} z' & \frac{1}{\rho} \end{bmatrix}, \quad (12)$$

where $z = (N_1^0)^{-1}u$, $\rho = \gamma - u'z$. The recursive algorithm, formulated by Eq. (12), will make the computation for N_1^{-1} easier, in comparison with the direct computation for the inverse matrix of N_1. Then $N_1^{-1}M$ and its summation of nonzero eigenvalues, $\lambda_{sum}(x)$, are calculated. After the above process has been done for each sample x, the sample corresponding to the maximum $\lambda_{sum}(x)$, denoted by Λ_s, is selected as the sth "significant node". The corresponding matrix $N_1^{-1}M$ is recorded and denoted by Q_s.

3.1.3. The termination on selecting "significant nodes"

Selecting for "significant node" is not terminated until $|\Lambda_s - \Lambda_{s-1}| < \varepsilon$, where ε is a constant. Suppose the number of the "significant nodes" is r, correspondingly the "significant nodes" are denoted by $x_1^o, x_2^o, \ldots, x_r^o$, respectively. All the eigenvectors $\alpha^1, \alpha^2, \ldots, \alpha^p$, corresponding to the nonzero eigenvalues of Q_r (suppose that there are p nonzero eigenvalues), must be solved.

3.2. Classification based on "significant nodes"

After "significant nodes" are found out, classification for test samples can be carried out based on them. For a test sample x_t, $f_m^o(x_t)$ is defined as

$$f_m^o(x_t) = \sum_{i=1}^{r} \alpha_i^m k(x_t, x_i^o), \quad m = 1, 2, \ldots, p, \quad (13)$$

where α_i^m is the ith component of eigenvector α^m. Furthermore, we define f_{im}^o as

$$f_{im}^o = \frac{1}{l_i} \sum_{j=1}^{l_i} \sum_{n=1}^{r} \alpha_n^m k(x_n^o, x_j^i),$$

$$i = 1, 2, \ldots, L, \; m = 1, 2, \ldots, p \quad (14)$$

and define the following vectors $F = [f_1^o(x_t) \; f_2^o(x_t) \; \cdots \; f_p^o(x_t)]$, $F_i^o = [f_{i1}^o \; f_{i2}^o \; \cdots \; f_{ip}^o]$, $i = 1, 2, \ldots, L$. The minimum distance classifier is adopted and classifying is performed based on the distances between F and every F_i^o, i.e., if F_j^o is the nearest to F, then x_t is classified into the jth class.

The algorithm presented in this section is suitable for both two-class problems and multi-class problems with more than two classes. On the other hand, the reformative algorithm for kernel Fisher discriminant analysis proposed in Ref. [2] is only for two-class problems.

Table 1
Experiment result on Yale Face Database

	Naïve kernel Fisher discriminant analysis	The reformative algorithm
Erroneous classification rate	12.2%	11.1%
The number of training samples or "significant nodes"	75	46(61%)

4. Experiments

In the following experiments, the naïve Fisher discriminant analysis also carries out classification using the minimum distance classifier. Different from the reformative algorithm, in the naïve Fisher discriminant analysis one test sample is classified based on all the kernel functions between the total training samples and the test sample.

4.1. Experiment result on Yale Face Database

Yale Face Database consists of 15 subjects. Each subject corresponds to 11 gray images, and different images vary much in expression and lighting condition. The first five images of each subject are taken as training samples, while the others are taken as test samples. Every image is treated as a vector and Gaussian kernel in the form of $k(x, y) = \exp(-\|x - y\|^2/(2\sigma^2))$ is adopted, σ^2 is set 1.0e + 8. In the experiment and the next one μ is set 0.001. Table 1 indicates that the erroneous classification rates achieved by the naïve kernel Fisher discriminant analysis and the reformative algorithm are 12.2% and 11.1%, respectively, while the reformative algorithm classifiers test samples only based on 46 "significant nodes", 61% of the total training samples. In other words, in the classification stage the computational time of the reformative algorithm is much less than the naïve kernel Fisher discriminant analysis.

4.2. Experiment on CMU mask images

CMU-Pittsburgh AU-Coded Face Expression Database [4] is a data set including happy, joy, angry, sad, surprise and disgusted expression face images. By using spatial adaptive triangulation technique based on local Gabor filters [5], 463 facial expression mask images are obtained. The first 210 mask images in the CMU database are taken as training samples, while the others are for testing. The solution of each image is 60×70. In the experiment kernel function $k(x_i, x_j) = (x_i \cdot x_j)^2$ is adopted and each image is also treated as a vector. Each gray-value is divided by 6000, so that matrices M and N corresponds to low condition numbers.

Table 2
Experiment result on CMU Face Expression Database

	Naïve kernel Fisher discriminant analysis	The reformative algorithm
Erroneous classification rate	11.5%	12.3%
The number of training samples or "significant nodes"	210	62(29.5%)

The experiment result given in Table 2 shows that the erroneous classification rates achieved by the naïve kernel Fisher discriminant analysis and the reformative algorithm are 11.5% and 12.3%, respectively, although the reformative algorithm classifiers test samples only based on 62 "significant nodes", 29.5% of the total training samples.

References

[1] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, K.-R. Müller, Fisher discriminant analysis with kernels, Neural Networks for Signal Processing IX, IEEE, 1999, pp. 41–48.
[2] Y. Xu, J-y Yang, J. Yang, A reformative kernel Fisher discriminant analysis, Pattern Recognition, accepted for publication.
[3] S.A. Billings, K.L. Lee, Nonlinear Fisher discriminant analysis using a minimum squared error cost function and the orthogonal least squares algorithm, Neural Networks 15 (2) (2002) 263–270.
[4] T. Kanade, J.F. Cohn, Y. Tian, Comprehensive database for facial expression analysis, in: Proceeding of the Fourth International Conference of Face and Gesture Recognition, Grenoble, France, 2000, pp. 46–53.
[5] S. Dubuisson, F. Davoine, M. Masson, A solution for facial expression representation and recognition, Signal Process. Image Commun. 17 (9) (2002) 657–673.

About the Author—YONG XU was born in Sichuan, China, in 1972. He received his B.S. degree and M.S. degree in 1994 and 1997, respectively. Now, he is working for his Ph.D. degree in Pattern recognition and Intelligence System. His current interests include face recognition and detection, character recognition and image processing.

About the Author—JING-YU YANG received the B.S. degree in Computer Science from NUST, Nanjing, China. His current research interests are in the areas of pattern recognition, image processing and artificial intelligence, and expert system.

About the Author—JIANFENG LU received his B.S., M.S., and Ph.D. degrees in Computer Science from Nanjing University of Science and Technology in 1991,1994,2000, respectively. Now he is a associate professor in NUST. His current research interests includes image processing, pattern recognition and computer vision.

About the Author—DONG-JUN YU was born in Jiangsu, China, on 19 October, 1975. In June 2003, he was granted Ph.D. degree in pattern recognition and intelligence system at Nanjing University of Science and Technology (NUST), Nanjing, China. Now he is an assistant at NUST.