PERGAMON

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

Pattern Recognition 37 (2004) 1299–1302

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Rapid and Brief Communication

A reformative kernel Fisher discriminant analysis

Yong Xu*, Jing-yu Yang, Jian Yang

Department of Computer Science, Nanjing University of Science and Technology, Nanjing Jiangsu 210094, People's Republic of China

Received 2 October 2003; accepted 15 October 2003

Abstract

A reformative kernel Fisher discriminant method is proposed, which is directly derived from the naive kernel Fisher discriminant analysis with superiority in classification efficiency. In the novel method only a part of training patterns, called "significant nodes", are necessary to be adopted in classifying one test pattern. A recursive algorithm for selecting "significant nodes", which is the key of the novel method, is presented in detail. The experiment on benchmarks shows that the novel method is effective and much efficient in classifying.

Keywords: Fisher discriminant analysis; Kernel trick; Pattern recognition

1. Introduction

Kernel Fisher discriminant analysis [1] has attracted much attention and has been applied in many recognition problems because it owns conceptual elegance and state-of-the-art performance. The key to kernel Fisher method is to classify test patterns in some high-dimensional space using kernel trick. However, it is well-known that kernel Fisher discriminant analysis is based on the theory of reproducing kernels; as a result, the classification efficiency of the naive kernel Fisher discriminant analysis is in verse ratio to the number of training patterns. Consequently, if the number of training patterns is large enough, the method may become impractical. So it is very important to improve the classification efficiency of the naive kernel Fisher discriminant analysis if a suitable approach is available [2,3]. In this paper, it is supposed that in the high-dimensional space introduced by kernel methods the Fisher discriminant vector can be well approximated by some expansion of a part of training patterns. From this supposition, a reformative kernel Fisher discriminant analysis is developed, which is directly derived from the naive kernel Fisher discriminant analysis and directly based on the Fisher criterion. Moreover, when the reformative method classifies each test pattern it is only necessary to compute the kernel functions between the test pattern and the "significant nodes", a few training patterns selected from the total training patterns. So, in practice, the novel method will be much more efficient than the naive kernel Fisher discriminant analysis in classifying.

2. Kernel Fisher discriminant analysis

Kernel Fisher discriminant analysis is based on a conceptual transformation from input space into a nonlinear high-dimensional feature space. Let $\{x_i\}$ denote the input space. Suppose that the high-dimensional 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

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

where w is the discriminant vector, and S_b^ϕ and S_w^ϕ are between-class scatter matrix and within-class scatter matrices, respectively. If there are two classes, they can be denoted by c_1 and c_2, respectively. The numbers of patterns in

* 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 © 2003 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2003.10.006

c_1 and c_2 are l_1 and l_2, respectively, meanwhile $l_1 + l_2 = l$ is supposed. $x_j^1,\ j = 1, 2, \dots, l_1$ denotes the jth pattern in c_1. $x_j^2,\ j = 1, 2, \dots, l_2$ denotes the jth pattern in c_2. If the prior probabilities of the two classes are equal, then

$$S_b^\phi = (m_1^\phi - m_2^\phi)(m_1^\phi - m_2^\phi)', \tag{2}$$

$$S_w^\phi = \sum_{i=1,2} \sum_{j=1,l_i} (\phi(x_j^i) - m_i^\phi)(\phi(x_j^i) - m_i^\phi)', \tag{3}$$

where $m_i^\phi = (1/l_i)\sum_{j=1,l_i} \phi(x_j^i)$. According to the theory of reproducing kernels, w will be an expansion of all training patterns, i.e.

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

Substitute $k(x_i, x_j)$ for dot production $\phi(x_i) \cdot \phi(x_j)$. Define M_1, M_2 and N as follows:

$$(M_i)_j = \frac{1}{l_i} \sum_{k=1}^{l_i} k(x_j, x_k^i), \qquad j = 1, 2, \dots, l, \tag{5}$$

$$N = \sum_{i=1}^{2} K_i(I - I_{l_i})K_i', \tag{6}$$

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, $(K_n)_{i,j} = k(x_i, x_j^n)$, $i = 1, 2, \dots, l,\ j = 1, 2, \dots, l_n,\ n = 1, 2$.

Fisher criterion based on kernel will be expressed by

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

where $\alpha = [\alpha_1 \dots \alpha_l]', \quad M = (M_1 - M_2)(M_1 - M_2)'$ [2].

As a result, the problem for obtaining w is transformed into one for solving optimal α, which corresponds to the maximum J. The optimal α will be solved by the eigenequation

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

So the eigenvector, corresponding to the maximum eigenvalue of Eq. (8), is the optimal α. In fact, for two-class classification the optimal α can be achieved by

$$\alpha = N^{-1}(M_1 - M_2). \tag{9}$$

3. The reformative kernel Fisher discriminant analysis

3.1. A reformative criterion function

In practice, generally N is singular, so α can be solved by the following equation:

$$\alpha = (N + \mu I)^{-1}(M_1 - M_2), \tag{10}$$

where μ is a positive constant. From the viewpoint of numerical stability, if μ is large enough, $N + \mu I$ will be positive definite and consequently the problem will be more stable. Now we define Fisher criterion as

$$J(\alpha) = \frac{\alpha' M \alpha}{\alpha' N \alpha + \mu \alpha' \alpha} = \frac{\alpha' M \alpha}{\alpha'(N + \mu I)\alpha}. \tag{11}$$

It is provable that

$$J(\alpha) = (M_1 - M_2)'\alpha. \tag{12}$$

Obviously, the greater $(M_1 - M_2)'\alpha$ is, the more significant the corresponding patterns are. So $(M_1 - M_2)'\alpha$ can be taken as a criterion to select "significant nodes". It is notable that although an algorithm based on criterion (11) favors α with small $\|\alpha\|_2$ it does not disobey Fisher's idea for achieving the maximum ratio of between-class distance to within-class distance.

3.2. Algorithm for selecting "significant nodes"

Step 1: *Selecting the first "significant node"*. For each training pattern $x_i,\ i = 1, 2, \dots, l$, first compute its N, M_1 and M_2 according to Eqs. (5) and (6). Then compute the corresponding α and $J(\alpha)$ according to Eqs. (10) and (12), respectively. The pattern corresponding to the maximum $J(\alpha)$ is taken as the first "significant node", denoted by x_1^0.

Step s: *Selecting the sth "significant node"*. Suppose $s-1$ patterns have been selected as "significant nodes", denoted by $x_1^o, x_2^o, \dots, x_{s-1}^o$, then selecting the sth "significant node" will be carried out according to the following algorithm.

It is notable that each pattern $x,\ x \in \{x_i,\ i = 1, 2, \dots, l\}$ and $x \notin \{x_j^0,\ j = 1, 2, \dots, s-1\}$, will be considered in this procedure. When a new pattern x is being considered, M_1, M_2, K_1, K_2 can be formulated as follows:

$$M_1 = \begin{bmatrix} M_1^0 \\ a \end{bmatrix}, \quad M_2 = \begin{bmatrix} M_2^0 \\ b \end{bmatrix}, \tag{13}$$

$$K_1 = \begin{bmatrix} K_1^0 \\ k_{new}^1 \end{bmatrix}, \quad K_2 = \begin{bmatrix} K_2^0 \\ k_{new}^2 \end{bmatrix}, \tag{14}$$

where M_1^0, M_2^0 are the M_1, M_2 corresponding to the previous $s-1$ "significant nodes", respectively, and K_1^0, K_2^0 are the K_1, K_2 corresponding to the previous $s-1$ "significant nodes", respectively,

$$a = \frac{1}{l_1} \sum_{k=1}^{l_1} k(x, x_k^1), \qquad b = \frac{1}{l_2} \sum_{k=1}^{l_2} k(x, x_k^2),$$

$$k_{new}^j = [k(x, x_1^j) \quad k(x, x_2^j) \quad \dots \quad k(x, x_{l_j}^j)], \qquad j = 1, 2.$$

Let

$$N_1 = \sum_{i=1,2} K_i(I - I_{l_i})K_i' + \mu I, \tag{15}$$

then N_1 can be expressed as

$$N_1 = \begin{bmatrix} N_1^0 & u \\ u' & \gamma \end{bmatrix}, \tag{16}$$

$$\gamma = \sum_{i=1,2} k_{new}^{i}(I - I_{l_i})(k_{new}^{i})' + \mu, \tag{17}$$

$$u = \sum_{i=1,2} K_i^0(I - I_{l_i})(k_{new}^{i})', \tag{18}$$

where N_1^0 is the N_1 corresponding to the previous $s-1$ "significant nodes". Because N_1 is a symmetric matrix, and N_1^{-1} can be obtained by the formulation [4]

$$N_1^{-1} = \begin{bmatrix} (N_1^0)^{-1} + \dfrac{1}{\rho} zz' & -\dfrac{1}{\rho} z \\ -\dfrac{1}{\rho} z' & \dfrac{1}{\rho} \end{bmatrix}, \tag{19}$$

where $z = (N_1^0)^{-1}u$, $\rho = \gamma - u'z$. Furthermore, $J(\alpha)$ can be calculated by the following:

$$J(\alpha) = (M_1^0 - M_2^0)'(N_1^0)^{-1}(M_1^0 - M_2^0) + [y - (a-b)]^2/\rho, \tag{20}$$

where $y = (M_1^0 - M_2^0)'z$. Because K_1^0, K_2^0, M_1^0, M_2^0, $(N_1^0)^{-1}$ and $(M_1^0 - M_2^0)'(N_1^0)^{-1}(M_1^0 - M_2^0)$ have been obtained in the procedure for selecting the $(s-1)$th "significant node", $J(\alpha)$ can be solved recursively and easily based on Eq. (20). After each pattern x has been considered and the corresponding $J(\alpha)$ has been obtained, the pattern corresponding to the maximum $J(\alpha)$, denoted by J_s, is selected as the sth "significant node". Selecting for the "significant node" is not terminated until $|J_s - J_{s-1}| < \varepsilon$, 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. α, corresponding to the maximum $J(\alpha)$, obtained in the procedure for selecting the last "significant node" is taken as the optimal solution for α.

3.3. Classification based on "significant nodes"

After "significant nodes" are selected classification for test patterns can be carried out based on them. For a test pattern x_t, $f(x_t)$ can be obtained by

$$f(x_t) = \sum_{i=1}^{r} \alpha_i k(x_t, x_i^o). \tag{21}$$

According to $f(x_t)$ classification can be performed. In the following experiment, the minimum classifier is exploited. In other words, if $f(x_t)$ is closer to f_1, x_t will be sorted into c_1, otherwise it will be sorted into c_2, where f_1 and f_2 are defined by

$$f_i = \frac{1}{l_i} \sum_{j=1}^{l_i} \sum_{k=1}^{r} k(x_k^o, x_j^i), \quad i = 1, 2. \tag{22}$$

4. Experiment

One experiment on 10 benchmark datasets (http://ida.first.gmd.de/∼raetsch/data/) is performed. Hundred partitions are generated for each dataset (except "Image" and "Splice" with only 20 partitions) and every partition includes own training pattern subset and test pattern subset. Gaussian kernel in the form of $k(x,y) = \exp(-\|x-y\|^2/(2\sigma^2))$ is adopted. For each dataset σ^2 is set the variance of the first training pattern subset. Training is also carried out in the first training subset, while classification is performed for all the test pattern subsets. In the experiment μ is set 0.001.

Tables 1 and 2 show the classification performance of the naive kernel Fisher discriminant analysis and the reformative method on 10 datasets. It is clear that classification error rates achieved by the reformative method are close to the naive kernel Fisher discriminant analysis. However, the number of "significant nodes" selected by the reformative method is much smaller than the total number of training patterns. The maximum ratio for "significant nodes" to the total training patterns is only 16%. In other words, only kernel functions between a few training patterns and one test pattern are used for classifying the test pattern in the reformative method.

Table 1
The classification error rates of the kernel Fisher discriminant analysis

Banana	B. Cancer	Diabetis	German	Heart	Image	F. Solar	Splice	Thyroid	Titanic
13.7 ± 0.1	22.7 ± 4.4	22.1 ± 1.9	21.3 ± 2.1	11.5 ± 2.8	9.0 ± 0.5	32.2 ± 1.6	11.0 ± 0.5	1.8 ± 1.1	25.5 ± 0.3

Table 2
The classification results of the reformative kernel Fisher discriminant analysis

	Banana	B. Cancer	Diabetis	German	Heart	Image	F. Solar	Splice	Thyroid	Titanic
Error rate	**13.3±0.1**	24.7 ± 4.1	23.5 ± 1.9	26.2 ± 2.0	**10.8±2.6**	11.4 ± 0.6	33.3 ± 1.6	13.7 ± 0.4	4.11 ± 1.8	**25.4±0.3**
"Significant nodes"	62(16%)	32(16%)	20(4%)	24(3%)	27(16%)	60(5%)	15(2%)	62(6%)	23(16%)	3(2%)

Consequently, it means that the computational complexity of classifying is much lower than that of the naive kernel Fisher discriminant analysis. Consequently, the reformative method is superior to the naive kernel Fisher discriminant analysis in classification efficiency.

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, New York, 1999, pp. 41–48.

[2] G.C. Cawley, N.L.C. Talbot, Efficient leave-one-out cross-validation of kernel Fisher discriminant classifiers, Pattern Recognition 36 (11) (2003) 2585–2592.

[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] G.H. Golub, C.F. van Loan, Matrix Computations, 3rd Edition, John Hopkins University Press, Baltimore, London, 1996.

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. He is the author of 8 scientific papers in pattern recognition and image processing. His current interests include face recognition and detection, character recognition and image processing.

About the Author—JING-YU YANG received his B.S. degree in Computer Science from NUST, Nanjing, China. He is the author of over 100 scientific papers in computer vision, pattern recognition, and artificial intelligence. His current research interests are in the areas of pattern recognition, image processing and artificial intelligence, and expert system.

About the Author—JIAN YANG was born in Jiangsu, China, in 1973. He obtained his B.S. degree, M.S. degree and Ph.D. degree in 1995, 1998 and 2002, respectively. Now, he is a postdoctor at the University of Zaragoza (Spain) and NUST (China). He is the author of more than 20 scientific papers in pattern recognition and computer vision. His current research interests include face recognition and detection, handwritten character recognition and data fusion.