Determine the Kernel Parameter of KFDA Using a Minimum Search Algorithm

Yong Xu[1], Chuancai Liu[2], and Chongyang Zhang[2]

[1] Department of Computer Science & Technology, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China
[2] Department of Computer Science & Technology, Nanjing University of Science & Technology, Nanjing, China
laterfall@hitsz.edu.cn, chuancailiu@yahoo.com.cn, zcy603@163.com

Abstract. In this paper, we develop a novel approach to perform kernel parameter selection for Kernel Fisher discriminant analysis (KFDA) based on the viewpoint that optimal kernel parameter is associated with the maximum linear separability of samples in the feature space. This makes our approach for selecting kernel parameter of KFDA completely comply with the essence of KFDA. Indeed, this paper is the first paper to determine the kernel parameter of KFDA using a search algorithm. Our approach proposed in this paper firstly constructs an objective function whose minimum is exactly equivalent to the maximum of linear separability. Then the approach exploits a minimum search algorithm to determine the optimal kernel parameter of KFDA. The convergence properties of the search algorithm allow our approach to work well. The algorithm is also simple and not computationally complex. Experimental results illustrate the effectiveness of our approach.

Keywords: Kernel Fisher discriminant analysis (KFDA); parameter selection; Linear separability.

1 Introduction

Kernel Fihser discriminant analysis (KFDA) [1-7] is a well-known and widely used kernel method. This method roots in Fisher discriminant analysis (FDA) [8-11]. FDA aims at achieving the optimal discriminant direction that is associated with the best linear separability. We can say that two procedures are implicitly contained in the implementation of KFDA. The first procedure maps the original sample space i.e. input space into a new space i.e. feature space and the second procedure carries out FDA in the feature space. Note that the feature space induced by KFDA is usually equivalent to a space obtained through a nonlinear transform. As a result, KFDA might produce linear-separable features for such data that are from the input space and have bad linear separability. On the other hand, FDA is not capable of doing so.

A kernel function is associated with KFDA and the parameter in the function is called kernel parameter. When we carry out KFDA, we should specify the value of the

D.-S. Huang, L. Heutte, and M. Loog (Eds.): ICIC 2007, LNAI 4682, pp. 418–426, 2007.

kernel parameter. Because different parameter values usually produce different feature extraction performances, to select a suitable value for the kernel parameter is significant. An expectation maximization algorithm developed by T. P. Centeno et al. determined the kernel parameter and the regularization coefficient through the maximization of the margin- likelihood of data [12]. Note that the optimization procedure in [12] is not guaranteed to find the global minimum. S. Ali and K. A. Smith have proposed an automatic parameter learning approach using Bayes inference [13]. The cross-validation criterion was also used to select free parameters in KFDA [14]. Though a nonlinear programming algorithm [15] can be applied to determine kernel and weighting parameters of a support vector machine, the effect depends on the choice of initial values of parameters. The DOE (design of experiments) technique was also used to select parameters for SVM machines [16]. These parameter selection approaches can be classified into two classes. The first class of approach usually determins the parameter value by maximizing the likelihood and the second class of approach is based on a criterion with respective to the relation between samples.

We can consider that the kernel parameter that results in the largest Fisher criterion is the optimal parameter. The rationale is as follows: first, the larger the Fisher criterion is, the greater linear separability different classes in the feature space have. Second, greater linear separability may allow higher classification performance to be produced. With this paper, we develop a novel kernel parameter selection approach for KFDA. This approach takes the maximization of Fisher criterion value as the target of parameter selection and uses a search algorithm. As far as the knowledge of the authors, no any other researcher has proposed the same parameter selection idea as ours. The theoretic property of the search algorithm can guarantee that the parameter selection approach has good performance. The moderate computation complexity allows parameter selection to be implemented efficiently. Moreover, the developed parameter approach does obtain good experimental results and gains performance improvement for KFDA. The other parts of this paper are organized as follows: KFDA are introduced briefly in Section 2. The idea and the algorithm of parameter selection are presented in Section 3. Experimental results are shown in Section 4. In Section 5 we offer our conclusion.

2 KFDA

KFDA [1], [2], can be derived formally from FDA as follows. Let $\{x_i\}$ denote the samples in the input space and let ϕ be a nonlinear function that transforms the input space into the feature space. Consequently, Fisher criterion in the feature space is

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

where w is a discriminant vector, S_b^{ϕ} and S_w^{ϕ} are respectively between-class and within-class scatter matrixes in the feature space. Suppose that there are two classes,

c_1 and c_2, and the numbers of samples in c_1 and c_2 are N_1 and N_2, respectively. Then the total number of the samples is $N = N_1 + N_2$. x_j^1, $j = 1,2,...,N_1$ denotes the $j-th$ sample in c_1. x_j^2, $j = 1,2,...,N_2$ means the $j-th$ sample in c_2. If the prior probabilities of the two classes are equal, then we have,

$$S_b^\phi = \left(m_1^\phi - m_2^\phi\right)\left(m_1^\phi - m_2^\phi\right)^T \tag{2}$$

$$S_w^\phi = \sum_{i=1,2} \sum_{j=1,2,...,N_i} \left(\phi\left(x_j^i\right) - m_i^\phi\right)\left(\phi\left(x_j^i\right) - m_i^\phi\right)^T \tag{3}$$

where $m_i^\phi = \frac{1}{N_i} \sum_{j=1,N_i} \phi\left(x_j^i\right), i = 1,2$. According to the theory of reproducing kernels, w can be an expressed in terms of all the training samples, i.e.

$$w = \sum_{i=1}^{N} \alpha_i \phi\left(x_i\right) \tag{4}$$

where each α_i is a scalar. We introduce a kernel function $k\left(x_i, x_j\right)$ to denote the dot production $\phi\left(x_i\right) \cdot \phi\left(x_j\right)$ and define M_1, M_2 and Q as follows:

$$\left(M_i\right)_j = \frac{1}{N_i} \sum_{s=1}^{N_i} k\left(x_j, x_s^i\right), j = 1,2,...,N, i = 1,2 \tag{5}$$

$$Q = \sum_{i=1}^{2} K_i \left(I - I_{N_I}\right) K_i^T \tag{6}$$

where I is the identity, I_{N_I} is an $N_i \times N_i$ matrix whose each element is $1/N_i$, K_i is a $N \times N_i$ matrix, $\left(K_n\right)_{i,j} = k\left(x_i, x_j^n\right), i = 1,2,...,N, j = 1,2,...,N_n, n = 1,2$.

Then we introduce notation M to mean the following formula:

$$M = \left(M_1 - M_2\right)\left(M_1 - M_2\right)^T \tag{7}$$

Note that Fisher criterion in the feature space can be expressed in terms of

$$J(\alpha) = \frac{\alpha^T M \alpha}{\alpha^T Q \alpha} \tag{8}$$

where $\alpha = [\alpha_1 \quad . \quad . \quad . \alpha_N]^T$.

As a result, the problem for obtaining the optimal discriminant vector w in the feature space can be converted into the problem for solving optimal α , which is associated with the maximum $J(\alpha)$. On the other hand, the optimal α will be obtained by solving the following eigenequation

$$M\alpha = \lambda Q\alpha . \tag{9}$$

After α is obtained, we can use it to extract features for samples. For detail please see [7]. Because the method presented above is defined on the basis of kernel function and Fisher discriminant analysis, it is called kernel Fisher discriminant analysis (KFDA). Note that the use of the kernel function allows KFDA to have a much lower computational complexity than an ordinary nonlinear Fisher discriminant analysis that implements explicitly FDA in the feature space obtained using a real mapping procedure. In addition, KFDA is able to obtain linear separable features for non-linear separable data whereas FDA cannot do so.

3 Select the Parameter Using a Search Algorithm

3.1 General Description of the Parameter Selection Scheme

As indicated in the context above, large Fisher criterion value means that the feature space has greater linear separability and higher classification accuracy can be expected. On the other hand, different parameter values of the kernel function will produce different Fisher criterion values. Consequently, the maximization of Fisher criterion (8) can be regarded as the objective of parameter selection. Note that the maximum of (8) coincides with the minimum of the following formual

$$J_2(\alpha) = \frac{\alpha^T Q \alpha}{\alpha^T M \alpha} \tag{20}$$

Thus, if a kernel parameter corresponds to a α that results in the minimum of (10), then the kernel parameter is the optimal parameter. In practice, if the M, Q associated with different kernel parameters are known, then the kernel parameter that is able to result in the minimum of (10) can be taken as the optimal parameter.

The Nelder-Mead simplex algorithm [17] is an enormously popular search algorithm for unconstrained minimization and it usually performs well in practice. The convergence properties of this search algorithm has been studied[18]. In fact, J. G. Lagarias has proved clearly that the algorithm converges to a minimizer for dimension 1. Moreover, the search algorithm is simple and not computationally complex.

3.2 Procedure of Parameter Selection

The following procedure can carry out the parameter selection scheme described in subsection 3.1:

Step 1. Set an initial value for the kernel parameter

Step 2. Calculate M, Q using (5), (6) and (7)

Step 3. Solving the smallest eigenvalue of $Q\alpha = \lambda M\alpha$.

Note that step 2 and step 3 will be repeatedly performed and will not be terminated until the convergence occurs. What the search algorithm does is to lead the computation to the convergence and to obtain the optimal kernel parameter that results in the minimum of (10).

3.3 Introduction to the Nelder-Mead Simplex Algorithm

The Nelder-Mead algorithm [18] focuses on minimizing a real-valued function f(x) for x$\in R^n$. Four scalar parameters exist in this method: coefficients of reflection (ρ), expansion (χ), contraction (γ), and shrinkage (σ). These parameters satisfy

$$\rho > 0, \chi > 1, \chi > \rho, 0 < \gamma, \sigma < 1. \tag{31}$$

At the beginning of the k th iteration, a nondegenerate simplex Δ_k is given, along with its n + 1vertices, each of which is a point in R^n. Assume that iteration k begins by ordering and labeling these vertices as $x_1^{(k)}, x_2^{(k)}, ..., x_{n+1}^{(k)}$, such that $f_1^{(k)} \le f_2^{(k)} \le ... \le f_{n+1}^{(k)}$, where $f_i^{(k)} = f(x_i^{(k)})$.- Note that the kth iteration generates n + 1 vertices that define a different simplex for the next iteration, so that $\Delta_{k+1} \neq \Delta_k$. The result of each iteration must be either of the follows: (i) a single new vertex, i.e. the accepted point, which replaces x_{n+1} in the set of vertices for the next iteration. (ii) if a shrink is performed, a set of n new points that, together with x_1, form the simplex at the next iteration. The Nelder-Mead algorithm can be implemented by the following iteration procedure [18]:

Step 1 (order). Order the n + 1 vertices to satisfy $f(x_1) \le f(x_2) \le ... \le f(x_{n+1})$ using the tie-breaking rules given below.

Step 2 (reflection). Calculate the reflection point x_r using $x_r = \bar{x} + \rho(\bar{x} - x_{n+1}) = (1+\rho)\bar{x} - \rho x_{n+1}$, where $\bar{x}$ denotes the mean of all vertices except for x_{n+1}. If $f_1 \le f_r < f_n$, the reflected point x_r should be accepted and then the iteration is terminated.

Step 3 (expansion). If $f_r < f_1$, we compute the expansion point x_e using $x_e = \bar{x} + \gamma(x_r - \bar{x}) = (1+\rho\chi)\bar{x} - \rho\chi x_{n+1}$. Then $f_e = f(x_e)$. If $f_e < f_r$, x_e will be accepted and the iteration will be terminated; otherwise, x_r will be accepted and the iteration will be terminated.

Step 4 (contraction). If $f_r \geq f_n$, we conduct a contraction between $\bar{x}$ and the better of x_{n+1} and x_r as follows.

(i) If $f_n \leq f_r < f_{n+1}$, let $x_c = \bar{x} + \gamma(x_r - \bar{x}) = (1+\rho\gamma)\bar{x} - \rho\gamma x_{n+1}$ and $f_c = f(x_c)$. If $f_c \leq f_r$, we accept x_c and terminate the iteration; otherwise, go to step 5.

(ii) If $f_r \geq f_{n+1}$, let $x_c = \bar{x} - \gamma(\bar{x} - x_{n+1}) = (1-\gamma)\bar{x} + \gamma x_{n+1}$ and $f_c = f(x_c)$. If $f_c < f_{n+1}$, we accept x_c and terminate the iteration; otherwise, go to step 5.

Step 5 (shrinkage). Evaluate f at the n points $v_i = x_1 + \sigma(x_i - x_1), i = 2,3,...,n+1$. Then the vertices of the simplex at the next iteration will be $x_1, v_2, ..., v_{n+1}$. The following rules are the so-called tie-breaking rules, which assign to the new vertex the highest possible index consistent with the relation $f(x_1^{(k+1)}) < f(x_2^{(k+1)}) \leq ... \leq f(x_{n+1}^{(k+1)})$.

(i) Nonshrink ordering rule. When a nonshrink step occurs, we discard the worst vertex $x_{n+1}^{(k)}$. Then the accepted point created during iteration k, denoted by $v^{(k)}$ becomes a new vertex and takes position $j+1$ in the vertices of Δ_{k+1}, where $j = \max_{0 \leq l \leq n} \{l \mid f(v^{(k)}) < f(x_{l+1}^{(k)})\}$. All other vertices retain their relative ordering from iteration k.

(ii) Shrink ordering rule. If a shrink step occurs, the only vertex carried over from Δ_k to Δ_{k+1} is $x_1^{(k)}$. Only one tie-breaking rule is specified, for the case in which $x_1^{(k)}$ and one or more of the new points are tied as the best point: if $\min\{f(v_2^{(k)}), f(v_3^{(k)}),..., f(x_{n+1}^{(k+1)})\} = f(x_1^{(k)})$, then $x_1^{(k+1)} = x_1^{(k)}$. A notation change index k^* of iteration k is defined as the smallest index of a vertex that differs between iterations k and $k+1$. When Nelder-Mead algorithm terminates in step 2, $1 < k^* \leq n$; for termination in step 3, $k^* = 1$; for termination in step 4, $1 \leq k^* \leq n+1$; and for termination in step 5, k^* is 1 or 2.

4 Experiments

We conducted experiments on several benchmark datasets to compare naive KFDA and KFDA with the parameter selection scheme. The kernel function employed in KFDA is the Gaussian kernel function $k(x_i, x_j) = \exp(\| x_i - x_j \|^2 / \eta)$. The minimum distance classifier was used for classification. For naive KFDA, the kernel parameter η are respectively set to be the norm of the covariance matrix of the training samples and its three times. For KDA with the parameter selection scheme, η are also respectively initially set to be the two values. Since each dataset has 100 training subsets and testing subsets, we conducted training and testing for every couple of training subset and testing subset. That is, if training was performed on the first training subset, testing would be carried out for the first test subset, and so on. As a result, for one dataset, we obtained 100 classification error rates respectively associated with 100 subsets. Then the mean and the standard deviation of the error rates were calculated. Table 1 indicates characteristics of these datasets. Table 2 and Table 3 respectively show classification results of naive KFDA and the KFDA model obtained using our parameter selection approach. Note that using the parameter selection scheme, KFDA obtained lower classification error rates.

Table 1. Characteristics of the datasets

	Banana	Diabetis	Heart	thyroid
Dimension of the sample vector	2	8	13	5
Number of classes	2	2	2	2
Sample number of each training subset	400	468	170	140

Table 2. Mean and standard deviation of classification error rates of naive KFDA on the subsets of one dataset. The first and second percentages denote the mean and standard deviation of classification error rates, respectively (the second percentage is written in the bracket). $\eta = \text{var}$ means that η is set the norm of the covariance matrix of the training samples.

	Banana	Diabetis	Heart	thyroid
$\eta = \text{var}$	12.99%(0.7%)	30.45%(2.2%)	23.16%(4.0%)	5.28 (3.0%)
$\eta = 3 \cdot \text{var}$	12.96%(0.8%)	27.15%(2.3%)	23.14%(3.5%)	5.39%(2.4%)

Table 3. Mean and standard deviation of classification error rates of our approach on the subsets of one dataset. The first and second percentages denote the mean and standard deviation of classification error rates, respectively (the second percentage is written in the bracket). $\eta = \mathrm{var}$ means that the initial value of η is set the norm of the covariance matrix of the training samples.

	banana	Diabetis	Heart	thyroid
$\eta = \sigma$	11.35%(0.6%)	26.40%(2.1%)	20.86%(3.6%)	5.08 %(2.2%)
$\eta = 3\sigma$	12.33%(0.7%)	25.92%(1.9%)	18.94%(3.2%)	5.10%(2.6%)

5 Conclusion

Our kernel parameter selection approach, which relates the optimal kernel parameter selection issue of KFDA with the Fisher-criterion maximization issue, is perfectly subject to the nature of FDA. This makes our approach be distinctive from all other parameter selection approaches. Additionally, one can understand easily the underlying reasonableness and rationality of our approach.

The underlying principle of our parameter selection approach is as follows: the optimal parameter should produce the best linear separability that is associated with the largest Fisher criterion value. Based on the defined objective function, whose minimum coincides with the maximum of Fisher-criterion, the approach developed in this paper can determine effectively the optimal kernel parameter by using a minimum search algorithm. In fact, the evidenced convergence property of the minimum search algorithm provides theoretical reasonability and practical feasibility with the parameter selection approach. Moreover, the fact that the search algorithm is simple and not computationally complex allows our approach to be carried out efficiently. Experimental results show that our approach allows the performance of KFDA to be greatly improved.

Acknowledgements. This work was supported by Natural Science Foundation of China（No. 60602038）and Natural Science Foundation of Guangdong Province, China（No. 06300862）.

References

1. Mika, S., Rätsch, G., Weston, J., et al.: Fisher Discriminant Analysis with Kernels. In: Y H Hu, J Larsen, E Wilson, S Douglas eds. Neural Networks for Signal Processing IX, IEEE, (1999) 41-48
2. Muller, K.-R., Mika, S., Ratsch, G., Tsuda, K., Scholkopf, B.: An Introduction to Kernel-based Learning Algorithms. IEEE Trans. On Neural Network, 12(1) (2001) 181-201
3. Billings, S.A., Lee, K.L.: Nonlinear Fisher Discriminant Analysis Using a Minimum Square Error Cost Function and the Orthogonal Least Squares Algorithm. Neural Networks, 15(1) (2002) 263-270
4. Yang, J., Jin, Z.H., Yang, J.Y., Zhang, D., Frangi, A.F.: Essence of Kernel Fisher Discriminant: KPCA plus LDA. Pattern Recognition 37(10) (2004) 2097-2100

5. Xu, Y., Yang, J.-Y., Lu, J., Yu, D.J.: An Efficient Renovation on Kernel Fisher Discriminant Analysis and Face Recognition Experiments. Pattern Recognition, 37 (2004) 2091-2094
6. Xu, Y., Yang, J.-Y., Yang, J.:A Reformative Kernel Fisher Discriminant Analysis. Pattern Recognition, 37 (2004) 1299-1302
7. Xu,Y., Zhang, D., Jin, Z., Li, M., Yang J.-Y.: A Fast Kernel-based Nonlinear Discriminant Analysis for Multi-class Problems. Pattern Recognition 39(6) (2006) 1026-1033
8. Duda, R., Hart, P.: Pattern Classification and Scene Analysis. New York: Wiley (1973)
9. P. Belhumeur, J. Hespanha, D. Kriegman.: Eigenface vs. Fisherface: Recognition Using Class Specific Linear Projection, IEEE Trans. Pattern Anal. And Mach. Intelligence, vol. 19, no. 10 (1997) 711-720
10. Xu, Y., Yang, J.Y., Jin, Z.:Theory Analysis on FSLDA and ULDA. Pattern Recognition, 36(12) (2003) 3031-3033
11. Xu, Y., Yang, J.-Y., Jin, Z.: A Novel Method for Fisher Discriminant Analysis. Pattern Recognition, 37(2) (2004) 381-384
12. Tonatiuh Peña Centeno, Neil D,Lawrence.: Optimising Kernel Parameters and Regularisation Coefficients for Non-linear Discriminant Analysis, Journal of Machine Learning Research 7 (2006) 455–491
13. Shawkat, Ali., Kate, A. Smith.: Automatic Parameter Selection for Polynomial Kernel, Proceedings of the IEEE International Conference on Information Reuse and Integration, USA (2003) 243-249
14. Volker, Roth.: Outlier Detection with One-class Kernel Fisher Discriminants. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, Advances in Neural Information Processing Systems 17, Cambridge, MA, MIT Press (2005) 1169–1176
15. Schittkowski, K.: Optimal Parameter Selection in Support Vector Machines, Journal of Industrial and Management Optimization, Vol. 1, No. 4, (2005) 465-476
16. Carl, Staelin.: Parameter Selection for Support Vector Machines, Technical report, HP Laboratories Israel (2003)
17. McKinnon, K.I.M.: Convergence of the Nelder-Mead to a No Stationary Point[J].SIAM Journal Optimization, 9 (1998) 148-158
18. Lagarias, J. G., Reeds, J. A., Wright, M.H., et al.: Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions, SIAM Journal of Optimization, 9(1) (1998) 112-147