Journal of Mathematical Sciences: Advances and Applications
Volume 24, 2013, Pages 13-27

A NOVEL DISCRIMINANT CRITERION FOR DATA CLASSIFICATION

LI LI and JIANQIANG GAO

Department of Applied Mathematics
Nanjing University of Finance and Economics
Nanjing 210023
P. R. China
e-mail: lili880827@126.com

College of Computer and Information Engineering
Hohai University
Nanjing 210098
P. R. China
e-mail: jianqianggaohh@126.com

Abstract

A calculation formula of recognition rate deviation wave (RRDW) is introduced in this paper. Meanwhile, the recognition performance of regularized linear discriminant analysis (RLDA) and maximum scatter difference discriminant analysis (MSD) methods were compared by using different parameters in three UCI data sets. The experimental results show that the recognition performance of RLDA is outperforms MSD under appropriate parameters. In addition, for different K values of K-nearest neighbour classifier (K-NNC), the RRDW of MSD is smaller than RLDA. Therefore, in practical applications, RRDW should be considered as a stable method to handle recognition tasks.

2010 Mathematics Subject Classification: 62H30, 68T10.

Keywords and phrases: RRDW, parameter selection, classification rate, RLDA, MSD.

Received November 16, 2013

1. Introduction

Dimensionality reduction and classification of data are important in many applications of data mining, machine learning, and face recognition. Many methods have been proposed for this problem, such as principal component analysis (PCA) [1] and LDA [2, 3]. PCA is one of the most popular methods for dimensionality reduction in compression and recognition problem, which tries to find eigenvector of components of original data. LDA is one of the most popular linear projection methods for feature extraction, it aims to maximize between-class scatter and minimize within-class scatter, thus maximize the class discriminant. An intrinsic limitation of LDA is that its objective function requires one of the scatter matrices being nonsingular. LDA cannot solve small sample problems [4], in which the data dimensionality is larger than the sample size and then all scatter matrices are singular.

In order to overcome the problems caused by the singularity of the scatter matrices, several methods have been proposed [5-7]. Recently, kernel methods have been widely used for nonlinear extension of linear algorithms [8-11]. The original data space is transformed to a feature space by an implicit nonlinear mapping through kernel methods. As long as an algorithm can be formulated with inner product computations, without knowing the explicit representation of a nonlinear mapping, we can apply the algorithm in the transformed feature space, obtaining nonlinear extension of the original algorithm. Gao and Fan [12] presented WKDA/QR and WKDA/SVD algorithms, which are used weighted function and kernel function use QR decomposition and singular value decomposition to solve small sample problems. In addition, by taking advantage of the technology of fuzzy sets [13], some studies have been carried out for fuzzy pattern recognition [14-18]. However, kernel-based and fuzzy methods are very time-consuming for computational complexities.

In addition, a maximum scatter difference discriminant analysis method (MSD) [19] was proposed by Song, which adopts the difference of both between-class scatter and within-class scatter as discriminant criterion, due to the inverse matrix is need not constructed, so the small sample problems occurred in LDA is avoided in nature.

In this paper, the recognition performances of RLDA and MSD are compared with respect to different parameters in three UCI [20] data sets. Meanwhile, the influence of K values also will be studied from K neighbour classifier. The results show that the comparison is very significantly.

The rest of the paper is organized as follows: The review of RLDA method, existing MSD model, and K-nearest neighbour classifier are briefly introduced and discussed in Sections 2, 3, and 4, respectively. A new calculation formula of recognition rate deviation wave (RRDW) is proposed in Section 5. In Section 6, experiments and analysis are reported. Conclusion follows in Section 7.

2. RLDA

Some important notations are introduced in this section. For matrix $A \in \Re^{n\times N}$, considering seek a linear transformation $G \in \Re^{n\times l}$ that maps each a_i of A to l-dimensional space $y_i \in \Re^l$ with $y_i = G^T x_i$. Assume that the original data in A is partitioned into c classes as $A = [A_1, A_2, \cdots, A_c]$, where $A_i \in \Re^{n\times N_i}$ contains sample data points from the i-th class and $N = \sum_{i=1}^{c} N_i$. Finding optimal transformation matrix G is a core problem of LDA model, and then the class structure of the original high-dimensional space is preserved in the reduced-dimensional space.

In discriminant analysis, the between-class, within-class, and total scatter matrices are defined as follows (see [3, 12, and 16]):

$$S_b = \frac{1}{N}\sum_{i=1}^{c} N_i (m_i - m)(m_i - m)^T = H_b H_b^T, \tag{1}$$

$$S_w = \frac{1}{N}\sum_{j=1}^{N_i}(m_i^j - m_i)(m_i^j - m_i)^T = H_w H_w^T, \tag{2}$$

$$S_t = S_b + S_w, \tag{3}$$

where the precursors H_b and H_w of the between-class and the within-class scatter matrices in (1) and (2) are

$$H_b = \frac{1}{\sqrt{N}}[\sqrt{N_1}(m_1 - m), \cdots, \sqrt{N_c}(m_c - m)], \tag{4}$$

$$H_w = \frac{1}{\sqrt{N}}[A_1 - m_1 e_1^T, \cdots, A_c - m_c e_c^T], \tag{5}$$

$e_i = (1, \cdots, 1)^T \in \Re^{N_i}$. A_i is the data matrix of the i-th class, m_i is the center of the i-th class, and m is the total cancroid of the training sample set. It is worthwhile to note that the total scatter matrix S_t can be called covariance matrix in statistics. Common optimization criteria for LDA are formulated as the maximization problem of objective functions.

$$J_1(G) = \frac{\text{trace}(G^T S_b G)}{\text{trace}(G^T S_w G)},$$

$$J_2(G) = \text{trace}((G^T S_w G)^{-1}(G^T S_b G)),$$

$$J_3(G) = \frac{|G^T S_b G|}{|G^T S_w G|}. \tag{6}$$

It is well-known that when S_w is nonsingular; the transformation matrix G is obtained by the eigenvectors corresponding to the $r-1$ largest eigenvalues of

$$S_w^{-1} S_b g = \lambda g. \tag{7}$$

In the RLDA [21], when S_w is singular or ill-conditioned, a diagonal matrix αI with $\alpha > 0$ is added to S_w. Since S_w is symmetric positive semi-definite, $S_w + \alpha I$ is nonsingular with $\alpha > 0$. So, the algorithm is used to solve the eigenvalues problem

$$S_b g = \lambda(S_w + \alpha I)g. \tag{8}$$

Therefore, in this paper, we use different parameters $\alpha(0.01 \le \alpha \le 1)$, step set to 0.1.

3. MSD

From the LDA method, the samples can be separated easily if the ratio of the between-class scatter and the within-class scatter is maximized. In this paper, a maximum scatter difference based discriminant rule is defined as follows [19]:

$$\begin{aligned} J(W) &= w^T S_b w - \alpha \cdot w^T S_w w \\ &= w^T (S_b - \alpha \cdot S_w) w, \end{aligned} \tag{9}$$

where α is a nonnegative constant to balance S_b and S_w.

By the property of the extreme value of generalized Raleigh quotient, the optimal solutions set maximizing (9) are the eigenvectors $w_1, w_2, \cdots, w_k$, corresponding to the first k largest eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_k$, where $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_k$. So, the optimal projection matrix is formed by $W = [w_1, w_2, \cdots, w_k]$. According to the analysis above, for a given sample x, its features can be obtained by transformation $y = W^T x$. When comparing the maximum scatter difference criterion (MSD) with the LDA criterion, we can easily find that the former avoids calculation of the inverse within-class scatter, i.e., $S_w^{-1} S_b$ is substituted by $S_b - \alpha \cdot S_w$, this cannot only make computationally more efficient but also avoid the singular problem of the within-class scatter. In this paper, we use different parameters α $(0.01 \le \alpha < 1)$, step set to 0.1.

4. *K*-NNC

In this section, on the one hand, the basic idea of the nearest neighbour classifier is introduced. On the other hand, the basic idea of *K*-nearest neighbour classifier (*K*-NNC) is pointed out. The nearest neighbour rule is quite simple, but very computationally intensive. For the digit example database, each classification requires 60,000 distance calculations between 784 dimensional vectors (28 × 28 pixels). The nearest neighbour code was therefore written in the C programming language in order to speed up the MATLAB testing.

Assume that the training samples data X with class identification, as $X = \{(x_1, l_1), \cdots, (x_n, l_n)\}$. If there is minimum distance between test sample x and one sample x_i of training samples data, and then, the category of test sample x belongs to l_i.

K-NNC is a nonparametric approach for classification. It does not require the priori knowledge, such as priori probabilities and the conditional probabilities. It operates directly towards the samples and is categorized as an instance-based classification method.

5. RRDW

In this section, the concept of range is introduced to pattern recognition. Some application can be found in [22]. We have been made the appropriate changes for range. In order to assess the merits of the RLDA and MSD model, we establish a recognition rate deviation wave (RRDW) criterion. MAX_{rr} denotes maximum recognition rate, MIN_{rr} denotes minimum recognition rate. So, this calculation is as follows:

$$\mathrm{RRDW} = \frac{(\mathrm{MAX}_{rr} - \mathrm{MIN}_{rr})}{\mathrm{MAX}_{rr}} * 100. \tag{10}$$

6. Experiments and Analysis

The performance of RLDA and MSD algorithms are evaluated on UCI database [20], which, namely, Iris plants, dermatology, and wine recognition databases are used for experiments. All experiments are performed on a PC (2.40GHZ CPU, 2G RAM) with MATLAB 7.1.

In order to get reliable experimental results, the following experiments have been done by using 5-fold cross validation method. The K-nearest neighbour algorithm (K-NN) with K = 1, 3, 5, 7, 9 is used as a classifier for all data sets, respectively. Meanwhile, the parameter α is set as 0.01, 0.11, 0.21, 0.31, 0.41, 0.51, 0.61, 0.71, 0.81, and 0.91, respectively.

6.1. Experiments with Iris plants. This is the best known database to be found in the pattern recognition literature. The data sets contain 3 classes and each class contains 4 attributes. For RLDA, the classification rates corresponding to each α parameter is illustrated in Table 1.

Table 1. Classification rates of Iris database (under different parameters α, RLDA-method)

K	0.01	0.11	0.21	0.31	0.41	0.51	0.61	0.71	0.81	0.91
1	91.3	92.0	93.3	94.0	94.0	95.3	95.3	94.6	95.3	95.3
3	90.0	92.0	92.7	92.7	94.0	94.0	93.3	93.3	96.0	96.7
5	88.7	89.3	92.0	92.7	92.7	93.3	93.3	92.7	94.7	94.0
7	88.7	91.3	92.0	92.7	92.7	93.3	92.7	94.0	94.0	94.7
9	88.0	89.3	90.7	91.3	92.7	91.3	92.7	92.7	93.3	94.0

For RLDA, the classification rates corresponding to parameters K are illustrated in Figure 1 ($\alpha = 0.91$).

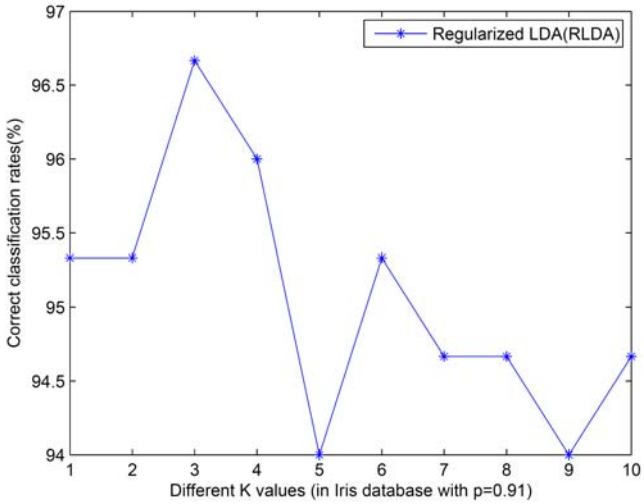

Figure 1. Classification rates of Iris database with different K.

For MSD, the classification rates corresponding to parameter α and K are illustrated in Figure 2.

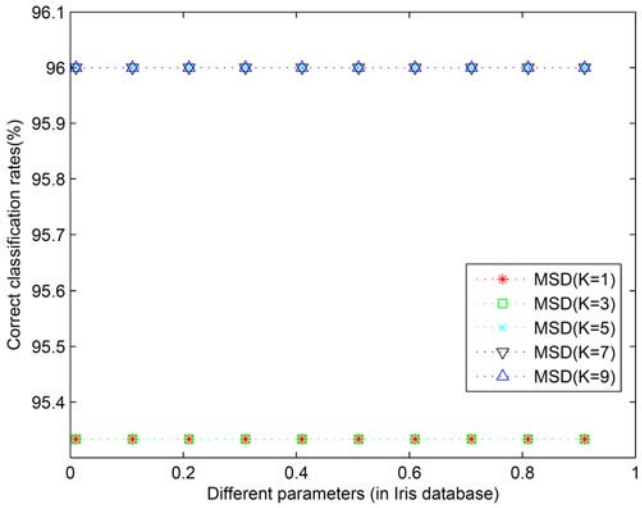

Figure 2. Classification rates of Iris database with different α.

From Table 1 and Figure 1, we can obtain the results as follows. On the one hand, the parameter α and K could play an important role in classification tasks. On the other hand, there are different changes of degree of dependence on parameter α. However, from Figure 2, we can see that the classification rates do not rely on any parameter α for MSD model. Therefore, from Figures 1 and 2, we can get the results by applying RRDW formula ($\alpha = 0.91$).

$$RRDW_{RLDA} = 2.1852\%,$$

$$RRDW_{MSD} = 0.7083\%. \tag{11}$$

According to the RRDW formula, MSD model is more stable than RLDA. However, RLDA is outperforms MSD model based on highest classification rate.

6.2. Experiments with dermatology. The data set contains 366 instances. Each instance contains 34 attributes. We have deleted the eight samples, which missing attribute values. The number of each class samples is different.

For RLDA, the classification rates corresponding to each α parameter is illustrated in Table 2.

Table 2. Classification rates of dermatology database with different parameters and K-NN classifier (RLDA)

K	0.01	0.11	0.21	0.31	0.41	0.51	0.61	0.71	0.81	0.91
1	95.6	94.5	94.5	93.9	93.5	93.6	93.9	95.2	95.5	95.3
3	96.4	95.9	95.3	95.8	95.3	95.3	96.7	97.2	95.6	96.1
5	96.7	96.4	96.2	96.4	96.1	96.4	96.7	96.6	96.1	96.8
7	97.3	96.7	96.5	96.1	96.7	97.3	96.4	96.9	96.1	95.6
9	97.0	97.5	96.5	96.4	97.3	97.5	96.5	96.7	96.7	96.2

For RLDA and MSD, the classification rates corresponding to parameters K are illustrated in Figures 3 and 4, respectively.

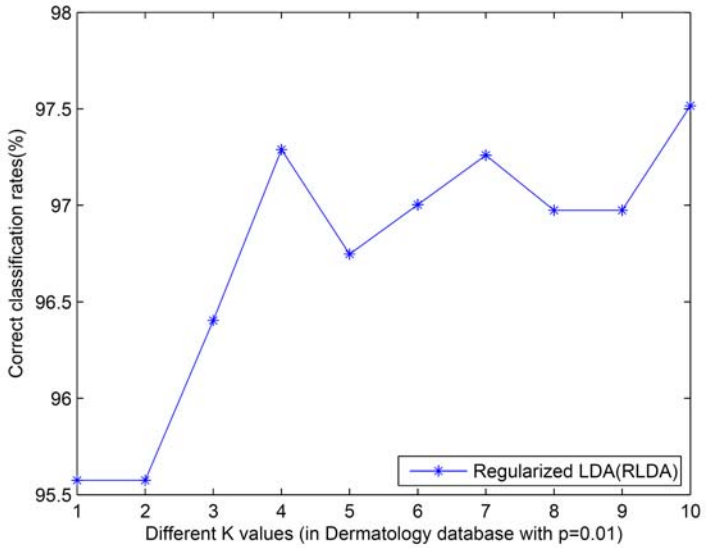

Figure 3. Classification rates of dermatology database with different K.

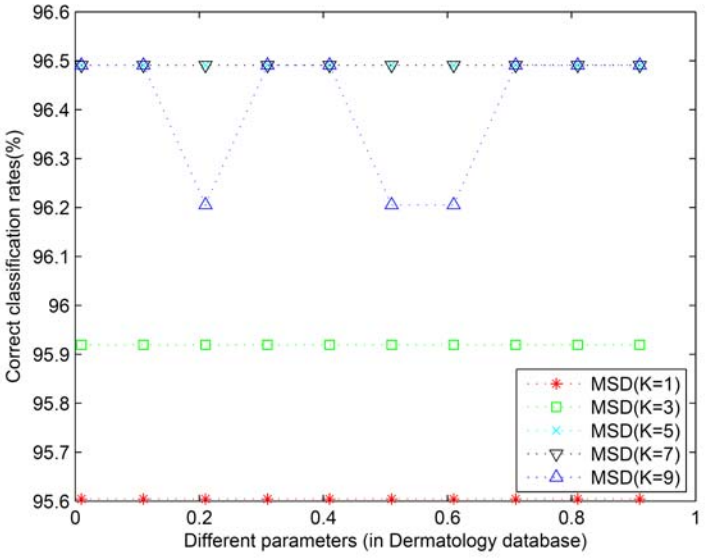

Figure 4. Classification rates of dermatology database with different α.

From Table 2 and Figure 3, we can obtain the same results like experiments with Iris database as follows. Firstly, the parameter α and K could play an important role in classification tasks. Secondly, there are different changes of degree of dependence on parameter α. But, in this data, small parameter α corresponds to the good results. However, from Figure 4, we can obtain that the classification rates do not rely on any parameter α for MSD model except for (K = 9). Meanwhile, the K value of K-NN classifier is very important for classification tasks. Therefore, from Figures 3 and 4, we get the results by using RRDW formula (α = 0.91).

$$RRDW_{RLDA} = 1.9588\%,$$

$$RRDW_{MSD} = 0.9326\%. \tag{12}$$

According to the RRDW, MSD model is more stable than RLDA. However, RLDA is outperforms MSD model according to highest classification rate.

6.3. Experiments with wine recognition. The data set contains 3 classes of 178 instances. Each class contains 13 attribute. For RLDA, the classification rates corresponding to each α parameter is illustrated in Table 3.

Table 3. Classification rates of wine recognition database with different parameters α

K	0.01	0.11	0.21	0.31	0.41	0.51	0.61	0.71	0.81	0.91
1	98.3	96.7	97.2	95.4	95.6	97.8	94.9	96.2	97.6	95.5
3	98.8	97.3	97.8	97.8	97.8	98.3	97.2	97.3	97.1	96.1
5	99.4	97.3	97.3	97.2	97.6	98.9	96.6	98.3	98.2	97.2
7	98.9	97.2	97.9	97.2	97.6	98.9	97.2	98.3	98.2	96.7
9	98.8	96.7	99.4	97.2	97.8	98.3	97.8	98.3	97.6	97.2

For RLDA and MSD, the classification rates corresponding to parameters K are illustrated in Figures 5 and 6, respectively.

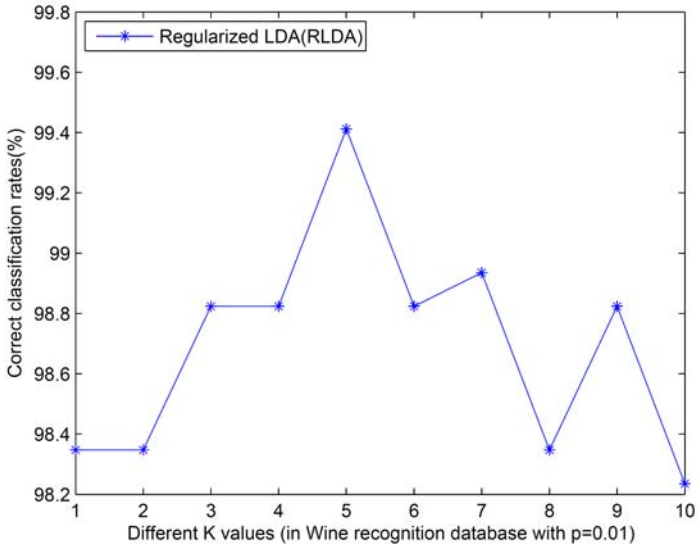

Figure 5. Classification rates of wine recognition database with different K.

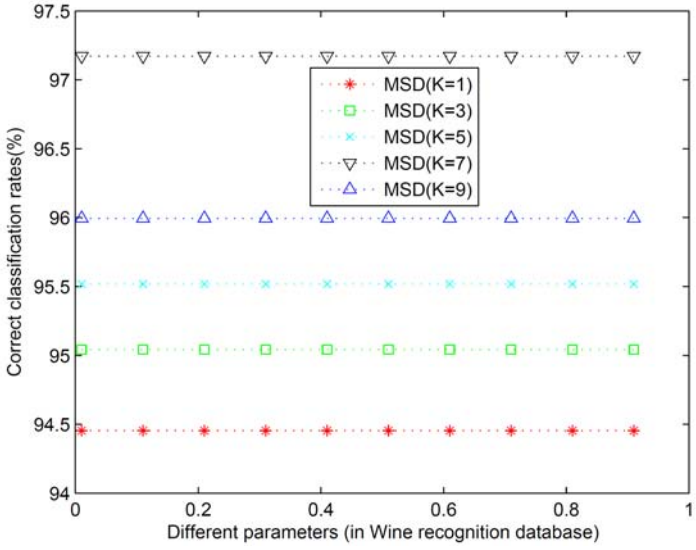

Figure 6. Classification rates of wine recognition database with different α.

As we can see from the Table 3 and Figure 5 that the importance of the parameter has aroused wide concern among the researcher, who care about the important of parameter optimization. But, for MSD model, the parameter α does not play a role in classification tasks. In addition, there is a quite important fact that the classification rates only rely on parameter K by using MSD model method. While $K = 9$, MSD model method has best classification effect. So, applying RRDW formula, we can get the results as follows ($\alpha = 0.91$):

$$RRDW_{RLDA} = 1.1972\%,$$

$$RRDW_{MSD} = 2.7881\%. \tag{13}$$

It is obvious that the recognition performance of RLDA model is outperforms MSD according to the RRDW and highest classification rate. That is to say the wine recognition data set has less redundant information.

7. Conclusion

In this paper, a novel discriminant criterion for data classification (RRDW) is proposed. Meanwhile, a comparison of regularized linear discriminant analysis and maximum scatter difference algorithms is also proposed for different parameters. We obtained different classification results by applying different methods. There are different advantages of RLDA and MSD methods. RLDA needs to compute inverse matrix, however, its classification results is better. And then, MSD does not need to compute inverse matrix, its computational complexity is small. Therefore, the comparison is very significantly.

We conduct the experiments on UCI data sets (Iris plants, dermatology, and wine recognition databases), using RLDA and MSD; the experimental results indicate that the parameter α and K value of K-NNC play an important role.

In Iris plants database, we should choose RLDA model and $K = 3$, in dermatology database, we also should choose RLDA model and $K = 10$, in wine recognition database, we also should choose RLDA model and $K = 5$.

However, for high dimensional data, in order to reduce computational complexity and obtain a stable method for recognition tasks, we should choose MSD model with more redundant information.

Through the above analysis, we can obtain following conclusions. The model and parameter are determined by applying RRDW formula. Therefore, the comparison of RLDA and MSD models are very significantly.

The future work on this subject will be investigating the influence of model parameter and kernel function in the classification problems. In addition, exploring the new algorithms to solve the corresponding optimization problems also is a further research direction.

Acknowledgement

The authors wish to thank the editor and the anonymous reviewers for a number of constructive comments and suggestions that have improved the quality of the paper. In addition, this work is supported by Graduate Education Innovation Project Fund of Jiangsu Province (CXZZ13-0239), P. R. China.

References

[1] I. T. Jolliffe, Principal Component Analysis, Second Edition, Springer-Verlag, New York, 2002.

[2] H. M. Lec, C. M. Chen and Y. L. Jou, An efficient fuzzy classifier with feature selection based on fuzzy entropy, IEEE Transactions on Systems, Man, and Cybernetics B 31(3) (2001), 426-432.

[3] J. Gao, L. Fan and L. Xu, Median null (Sw)-based method for face feature recognition, Applied Mathematics and Computation 219(12) (2013), 6410-6419.

[4] J. Ye, R. Janardan, C. H. Park and H. Park, An optimization criterion for generalized discriminant analysis on under sampled problem, IEEE Transaction, Pattern Analysis and Machine Intelligence 26(8) (2004), 982-994.

[5] J. W. Lu, K. N. Plataniotis and A. N. Venetsanopoulos, Regularized discriminant analysis for the small sample size problem in face recognition, Pattern Recognition Letters 24 (2003), 3079-3087.

[6] Y. Q. Guo, T. Hastie and R. Tibshirani, Regularized linear discriminant analysis and its application in micro-arrays, Biostatistics 8(1) (2007), 86-100.

[7] W. Zheng, L. Zhao and C. Zou, An efficient algorithm to solve the small sample size problem for LDA, Pattern Recognition 37 (2004), 1077-1079.

[8] J. Yang, Z. Jin, J. Yang and D. Zhang, The essence of kernel Fisher discriminant: KPCA plus LDA, Pattern Recognition 37 (2004), 2097-2100.

[9] J. Yang, A. F. Frangi, J. Y. Yang and D. Zhang, KPCA plus LDA: A complete kernel Fisher discriminant framework for feature extraction and recognition, IEEE Transaction, Pattern Anal. Mach. Intel. 27 (2005), 230-244.

[10] B. H. Shekar et al., Face recognition using kernel entropy component analysis, Neurocomputing 74 (2011), 1053-1057.

[11] A. Eftekhari et al., Block-wise 2D kernel PCA/LDA for face recognition, Information Processing Letters 110 (2010), 761-766.

[12] J. Gao and L. Fan, Kernel-based weighted discriminant analysis with QR decomposition and its application face recognition, WSEAS Transactions on Mathematics 10(10) (2011), 358-367.

[13] L. A. Zadeh, Fuzzy sets, Information Control 8 (1965), 338-353.

[14] K. C. Kwak and W. Pedrycz, Face recognition using a fuzzy Fisher face classifier, Pattern Recognition 38(10) (2005), 1717-1732.

[15] W. Li, Q. Ruan and J. Wan, Fuzzy nearest feature line-based manifold embedding, Journal of Information Science and Engineering 29 (2013), 329-346.

[16] Yan. Cui and L. Fan, Feature extraction using fuzzy maximum margin criterion, Neurocomputing 86 (2012), 52-58.

[17] W. An and M. Liang, Fuzzy support vector machine based on within-class scatter for classification problems with outliers or noises, Neurocomputing (2013),

http://dx.doi.org/10.1016/j.neucom.2012.11.023.

[18] J. Gao and L. Fan, The impact on the face recognition from distances in fuzzy linear discriminant analysis, Journal of Jinggang-shan University (Natural Science) (in Chinese) 33(3) (2012) 1-7. DOI:10.3969/j.issn.1674-8085.2012.03.001.

[19] F. X. Song, K. Cheng and J. Y. Yang et al., Maximum scatter difference, large margin linear projection and support vector machines, Acta Automatica Sinica 30(6) (2004), 890-896 (in Chinese).

[20] http://archive.ics.uci.edu/ml/datasets.html. Accessed 10, January 2013.

[21] W. Ching et al., Regularized orthogonal linear discriminant analysis, Pattern Recognition 45 (2012), 2719-2732.

[22] K. Fang and Z. Liu, The application of range in variance analysis, Mathematics in Practice and Theory 1 (1976), 37-51 (in Chinese).

■