A minimum distance-based method for the classification problem

Jianqiang Gao
Hohai University
College of Computer and Information
Jiangning 211100, Nanjing
P.R. China
Correspondence: jianqianggaohh@126.com

Lizhong Xu
Hohai University
College of Computer and Information
Jiangning 211100, Nanjing
P.R. China
mathtwo@hotmail.co.in

Abstract: In this paper, a kernel fuzzy discriminant analysis minimum distance-based approach for the classification of face images is proposed to deal with face classification problem (we call this method mdkfda/qr as an abbreviation). A superiority of the mdkfda/qr is its computational efficiency and can avoid the singularity. In the proposed method, the membership degree is incorporated into the definition of between-class and within-class scatter matrixes to get fuzzy between-class and within-class scatter matrixes. The mdkfda/qr approach was compared with kernel discriminant analysis (KDA) and fuzzy discriminant analysis (FDA) two algorithms in terms of classification accuracy. Experiments on ORL and FERET two real face datasets are performed to test and evaluate the effectiveness of the proposed algorithm on classification accuracy. The results show that the effect of mdkfda/qr method can achieve higher classification accuracy than KDA and FDA methods.

Key–Words: Kernel discriminant analysis, Fuzzy membership, QR decomposition, Classification, mdkfda/qr

1 Introduction

Linear discriminant analysis (LDA) is one of the most popular linear projection techniques for feature extraction. LDA seeking optimal linear projections such that the Fisher criterion of the between-class scatter versus the within-class scatter is maximized, also is a one of the most well-known statistical technique for feature extraction and dimension reduction [1]. Due to the singularity of within-class scatter matrix, it cannot be applied directly to small size sample problems [2]. In order to use LDA for small size sample problems such as face recognition, much research work has been obtained [3-8]. In addition, several extensions of LDA [9-13] have been developed concerning robustness issue. The most popular approach, Fisher face was build by D.L. Swets et al. [6] and Belhumeur et.al. [7], in which principal component analysis (PCA) is first used to reduce the dimension of the original (feature) space and then the classical Fisher linear discriminant analysis (FLDA) is applied in reduction dimension space. A limitation of Fisher face is that some effective discriminatory information may be lost and use PCA step cannot guarantee the transformed within-class scatter matrix be nonsingular.

Kernel-based learning methods have attracted much attention in the areas of pattern recognition and machine learning, such as support vector machines (SVM) [14], kernel principal component analysis (KPCA) [15], kernel canonical correlation analysis (KCCA) [16]. Mika et al. [17] proposed kernel discriminant analysis (KDA) for two-class cases. Yang et al. [18] further discussed kernel Fisher discriminant analysis and pointed out that kernel Fisher discriminant analysis is equivalent to kernel principal component analysis plus Fisher linear discriminant analysis. Hence, for high-dimensional multiclass tasks such as faces recognition, the original KDA-based algorithms usually encounter three difficulties: the first is the singularity problem caused by the small size sample problems. The second is that the Fisher separability criterion is not directly related to classification rate, that is to say the classes with larger distance to each other in feature space are more emphasized when the Fisher criterion is optimized, which leads that the resulting projection preserves the distance of already well separated classes, causing a large overlap of neighboring classes [4]. The third is that these algorithms still face the computational difficulty of the eigen-decomposition of matrices in the high-dimensional space. For the above mentioned three problems in many applications, it is necessary to develop new and more effective algorithm to deal with them.

Recently, Dai et al. [19-20], Zhou and Tang [21-22] presented kernel-weighted discriminant analysis by generalizing the fractional LDA [23]. The main methods in [21-22] are the simultaneous digitalization technique for tackling the small size sample problems. By taking advantage of the technology of fuzzy sets

[24], some studies have been carried out for fuzzy pattern recognition [25-26].

Inspired and motivated by above mentioned, we extend fuzzy Fisher discriminant analysis (FDA) with minimum distance obtain kernel fuzzy discriminant analysis approach mdkfda/qr, which based on QR decomposition. Since QR decomposition on a small size matrix is adopted, a superiority of our method is its computational efficiency and can avoid the singularity. In the proposed method, the membership degree is incorporated into the definition of between-class and within-class scatter matrixes to get fuzzy between-class and within-class scatter matrixes.

The rest of this paper is organized as follows. The KDA and FDA are briefly introduced and discussed in Section 2. The detailed description of mdkfda/qr approach is presented in Section 3. In Section 4, to demonstrate the effectiveness of our method, we compare mdkfda/qr with some known algorithms. Conclusions are summarized in Section 5.

2 Review of KDA and FDA

2.1. KDA

Kernel discriminant analysis (KDA) is a kernel version of LDA to deal with the feature extraction and classification of nonlinear characteristics. The basic idea of KDA is to firstly project original patterns into a high-dimensional feature space F by an implicit nonlinear mapping $\phi : \mathbb{R}^n \to \mathbb{F} : x \to \Phi(x)$ and then to use LDA in feature space $\mathbb{F}$.

Let us consider a set of m training samples $\{x_1, x_2, \cdots, x_m\}$ taking values in an n dimensional space. Let L be the number of classes and m_i the number of training samples in the i-th class, $i = 1, \cdots, L$. Obviously, $m = \sum_{i=1}^{L} m_i$. In general, the Fisher criterion [18] can be defined as

$$\max_{v} J(v) = \frac{v^T S_b^{\phi} v}{v^T S_t^{\phi} v}, \tag{1}$$

where $S_b^{\phi} = \frac{1}{m}\sum_{i=1}^{L} m_i(m_i^{\phi} - m_0^{\phi})(m_i^{\phi} - m_0^{\phi})^T$ and $S_t^{\phi} = \frac{1}{m}\sum_{i=1}^{m}(\phi(x_i) - m_0^{\phi})(\phi(x_i) - m_0^{\phi})^T$ are the between-class and total scatter matrixes defined in the feature space $\mathbb{F}$, respectively, where m_i^{ϕ} is the mean vector of the mapped training samples in the i-th class and m_0^{ϕ} is the mean vector of all mapped training samples. The optimization problem (1) can be transformed into the following eigenvalue problem:

$$S_b^{\phi} v = \lambda S_t^{\phi} v. \tag{2}$$

Let $\Phi(X) = [\phi(x_1), \cdots, \phi(x_m)]$ and $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be a kernel function. The kernel matrix $K = (k_{ij}) \in \mathbb{R}^{m \times m}$ corresponded to the kernel k can be defined by $k_{ij} = k(x_i, x_j) =< \phi(x_i), \phi(x_j) >$, where $\phi : \mathbb{R}^n \to \mathbb{F}$ is a feature map and $\mathbb{F}$ is a feature space of the kernel k. It is evident that $K = \Phi(X)^T\Phi(X)$. For any $j \in \{1, \cdots, m\}$, let $\widetilde{\phi}(x_j) = \phi(x_j) - \frac{1}{m}\sum_{i=1}^{m}\phi(x_i)$ be the centered mapped data and $\widetilde{\Phi}(X) = [\widetilde{\phi}(x_1), \cdots, \widetilde{\phi}(x_m)] = \Phi(X)(I - 1_{m\times m}/m)$, where I is a $m \times m$ identity matrix and $1_{m\times m}$ is a $m \times m$ matrix of all ones. The inner product matrix $\widetilde{K}$ for the centered mapped data can be obtained by

$$\begin{aligned} \widetilde{K} &= \widetilde{\Phi}(X)^T\widetilde{\Phi}(X) \\ &= (I - 1_{m\times m}/m)^T K (I - 1_{m\times m}/m). \end{aligned} \tag{3}$$

According to the reproducing kernel theory [15], the eigenvector v lies in the span of $\{\widetilde{\phi}(x_1), \cdots, \widetilde{\phi}(x_m)\}$ and then there exist coefficients $a_i, (i = 1, 2, \cdots, m)$ such that

$$v = \sum_{i=1}^{m} a_i\widetilde{\phi}(x_i) = \widetilde{\Phi}(X)a, \tag{4}$$

where $a = (a_1, \cdots, a_m)^T$. Let $W = diag(s_1, \cdots, s_j, \cdots, s_L)$, where s_j is a $m_j \times m_j$ matrix whose elements are $1/m_j$. Substituting (4) into (1), we can obtain the following equation:

$$\max_{a} J(a) = \frac{a^T\widetilde{K}W\widetilde{K}a}{a^T\widetilde{K}\widetilde{K}a}. \tag{5}$$

Generally speaking, the vector a_1 corresponding to the maximal value of $J(a)$ is the optimal discriminant direction. However, in some cases, it is not enough to only use one optimal discriminant direction to feature extraction. Hence, it is often necessary to obtain t $(t > 1)$ optimal discriminant directions. Assume that $a_1, \cdots, a_t$ are t optimal discriminant directions and $A = [a_1, a_2, \cdots, a_t]$. Then A should satisfy

$$A = \arg\max_{A} \mathrm{tr}\left(\frac{A^T S_b' A}{A^T S_t' A}\right), \tag{6}$$

where $S_b' = \widetilde{K}W\widetilde{K}$, $S_t' = \widetilde{K}\widetilde{K}$, and $tr(\cdot)$ denotes the trace of matrices. The optimization problem (6) can be transformed into the following generalized eigenvalue problems:

$$S_b' a = \lambda S_t' a. \tag{7}$$

The solution of Eq. (7) can be obtained by solving the generalized eigenvalue problem. Suppose that $\lambda_1, \lambda_2, \cdots, \lambda_t$ are the t largest eigenvalues of the Eq. (7) sorted in descending order and $a_1, \cdots, a_t$ are the

corresponding eigenvectors. The KDA transform matrix can obtained by using Eq. (8).

$$\begin{aligned} V = [v_1, \cdots, v_t] &= \widetilde{\Phi}(X)[a_1, \cdots, a_t] \\ &= \widetilde{\Phi}(X)A. \end{aligned} \tag{8}$$

For any input vector x, its low dimension feature representation y_x can be defined by

$$\begin{aligned} y_x &= V^T\widetilde{\phi}(x) \\ &= A^T\widetilde{\Phi}(X)^T\widetilde{\phi}(x) \\ &= A^T(\widetilde{k}(x_1, x), \cdots, \widetilde{k}(x_m, x))^T. \end{aligned} \tag{9}$$

2.2. FDA

In [26], Kwark et al. proposed the fuzzy fisher face approach for recognition by fuzzy set. Given a set of feature vectors. $X = \{x_1, x_2, \cdots, x_m\}$, L is known pattern classes $X_1, X_2, \cdots, X_L$. m_i is the number of training samples of class i and satisfies $m = \sum_{i=1}^{L} m_i$, $M_0 = (1/m)\sum_{i=1}^{L} x_i$. A fuzzy L-class partition of these vectors specifies the degree of membership of each vector to the classes. The membership matrix $[u_{ij}](i = 1, 2, \cdots, L, j = 1, 2, \cdots, m)$ can be got by FKNN [24]. Taking into account the membership grades, the mean vector of each class M_i is calculated as follows:

$$M_i = \frac{\sum_{j=1}^{m} u_{ij}x_j}{\sum_{j=1}^{m} u_{ij}}. \tag{10}$$

The between-class fuzzy scatter matrix S_{Fb} and within-class fuzzy scatter matrix S_{Fw} incorporate the membership values in their calculations.

$$S_{Fb} = \frac{1}{m}\sum_{i=1}^{L} m_i(M_i - M_0)(M_i - M_0)^T. \tag{11}$$

$$S_{Fw} = \frac{1}{m}\sum_{i=1}^{L}\sum_{j=1}^{m_i}(x_i^j - M_i)(x_i^j - M_i)^T. \tag{12}$$

The optimal fuzzy projection matrix W of fuzzy fisher face follows the expression:

$$W = arg\max_{W} \frac{|W^T S_{Fb} W|}{|W^T S_{Fw} W|}. \tag{13}$$

Finally, PCA plus fuzzy LDA is used in small size sample cases.

3 mdkfda/qr algorithm

FDA is linear learning approach and it cannot deal with nonlinear problem. In order to solve this problem, we introduce minimum distance-based weighted LDA and kernel skill into FDA to obtain mdkfda/qr. The main idea of mdkfda/qr is that original samples are projected firstly into a feature space of a kernel function by an implicit feature mapping and then use minimum distance-based weighted LDA, where the between-class fuzzy scatter matrix in the feature space is defined by pairwise weighted functions. In minimum distance-based weighted LDA, the QR decomposition is used to find low dimensional nonlinear feature with significant discrimination power, respectively.

3.1. Minimum distance

Let $X = [X_1, X_2, \cdots, X_p]^T$, which have p features, $x_i = [x_{i1}, x_{i2}, \cdots, x_{ip}]^T$, $(i = 1, 2, \cdots, n)$ are n samples. Every sample can be seen a point with p-dimension space. In here, let $d(x_i, x_j)$ is the distance between x_i and x_j, where x_i and x_j are samples. The definition of minimum distance as follows:

$$d(x_i, x_j) = \min_{1\le k\le p} |x_{ik} - x_{jk}|.$$

where $i = 1, 2, \cdots, n, k = 1, 2, \cdots, p$.

3.2. Fuzzy K-nearest neighbor (FKNN)

In our method, fuzzy membership degree and each class center are obtained through FKNN [24] algorithm. With FKNN algorithm, the computations of the membership degree can be realized through a sequence of steps:

Step 1: Compute the minimum distance matrix of feature vectors in training set.

Step 2: Set diagonal elements of this minimum distance matrix to infinity.

Step 3: Sort the distance matrix (treat each of its columns separately) in an ascending order. Collect the corresponding class labels of the patterns located in the closest neighborhood of the pattern under consideration (as we are concerned with "k" neighbors, this returns a list of "k" integers).

Step 4: Compute the membership degree to class "i" for j-th pattern using the expression proposed in the literature [24].

$$u_{ij} = \begin{cases} 0.51 + 0.49 \times (n_{ij}/k), \text{if } i\text{=the same as the} \\ \text{j-th label of the pattern.} \\ 0.49 \times (n_{ij}/k), \text{if } i \neq \text{the same as the} \\ \text{j-th label of the pattern.} \end{cases}$$

In the above expression n_{ij} stands for the number of the neighbors of the j-th patter that belong to the i-th class. As usual, u_{ij} satisfies two obvious properties: $\sum_{i=1}^{L} u_{ij} = 1, 0 < \sum_{j=1}^{m} u_{ij} < N$. Therefore, the fuzzy membership matrix U can be achieved with the help of FKNN. $U = [u_{ij}]$, $(i = 1, 2, \cdots, L, j = 1, 2, \cdots, m)$.

3.3. Weighted schemes

In order to obtain a modified criterion that it is more closely related to classification error, weighted schemes can be introduced into the traditional Fisher criterion to penalize the classes that are close in the feature space and then lead to potential misclassifications in the output space.

Let $d_0 = d^{\phi}(M_i^{\phi}, M_0^{\phi})$ be distance between the mean of class i and the mean of total. $w(\cdot)$ be a weighted function, which is usually a monotonically decreasing function. $D = d^{\phi}(\widetilde{M}_i^{\phi}, \widetilde{M}_0^{\phi})$ The weighted between-class fuzzy scatter of the centered samples in the feature space $\mathbb{F}$ can be defined as follows:

$$S_{Fb}^{\phi w} = \frac{1}{m}\sum_{i=1}^{L} m_i w(d_0)(\widetilde{M}_i^{\phi} - \widetilde{M}_0^{\phi})(\widetilde{M}_i^{\phi} - \widetilde{M}_0^{\phi})^T,$$

where $w(d_0) = (D)^{-q}$ and $q \geq 2$. If $w(\cdot) = 1$, the matrix $S_{Fb}^{\phi w}$ will degenerate to the matrix $S_{Fb}^{\phi} = \frac{1}{m}\sum_{i=1}^{L} m_i(M_i^{\phi} - M_0^{\phi})(M_i^{\phi} - M_0^{\phi})^T$. In this paper, we use the Euclidean distance. Since $\widetilde{M}_0^{\phi} = 0$ and $\widetilde{M}_i^{\phi} = \widetilde{\Phi}(X)U^T e_i$, where $e_i = [\underbrace{0,\cdots,0}_{0+\cdots+i-1}, \underbrace{1}_{i}, \underbrace{0,\cdots,0}_{L-i}]^T$, we have

$$\begin{aligned} d_0 &= d^{\phi}(\widetilde{M}_i^{\phi}, \widetilde{M}_0^{\phi}) = \sqrt{(\widetilde{M}_i^{\phi})^T(\widetilde{M}_i^{\phi})} \\ &= \sqrt{e_i{}^T U\widetilde{\Phi}(X)^T\widetilde{\Phi}(X)U^T e_i} \\ &= \sqrt{e_i{}^T U\widetilde{K}U^T e_i}. \end{aligned}$$

Putting $d_i = \sqrt{e_i{}^T U\widetilde{K}U^T e_i}$, we can deduce that

$$\begin{aligned} S_{Fb}^{\phi w} &= \frac{1}{m}\sum_{i=1}^{L} m_i w(d_0)(\widetilde{M}_i^{\phi})(\widetilde{M}_i^{\phi})^T \\ &= \frac{1}{m}\widetilde{\Phi}(X)U^T HU\widetilde{\Phi}(X)^T, \end{aligned}$$

$$\begin{aligned} S_{Ft}^{\phi} &= \frac{1}{m}\sum_{j=1}^{m}(\phi(x_j) - M_0^{\phi})(\phi(x_j) - M_0^{\phi})^T \\ &= \frac{1}{m}\widetilde{\Phi}(X)E\widetilde{\Phi}(X)^T \\ &= \frac{1}{m}\widetilde{\Phi}(X)\widetilde{\Phi}(X)^T, \end{aligned}$$

where $H = diag(m_1 d_1^{-q}, \cdots, m_j d_j^{-q}, \cdots, m_L d_L^{-q})$, $E_j = [\underbrace{0,\cdots,0}_{0+\cdots+j-1}, \underbrace{1}_{j}, \underbrace{0,\cdots,0}_{m-j}]^T$. Therefore, the optimal transform matrix V^{ϕ} can be obtained by maximizing the following Fisher criterion:

$$V^{\phi} = \arg\max_{V^{\phi}} \mathrm{tr}\left(\frac{{V^{\phi}}^T S_{Fb}^{\phi w} V^{\phi}}{{V^{\phi}}^T S_{Ft}^{\phi} V^{\phi}}\right). \qquad (14)$$

By means of the kernel trick, the optimization problem (14) can be transformed to the following optimization problem:

$$\widetilde{A} = \arg\max_{\widetilde{A}} \mathrm{tr}\left(\frac{\widetilde{A}^T SB\widetilde{A}}{\widetilde{A}^T ST\widetilde{A}}\right), \qquad (15)$$

where $V^{\phi} = \widetilde{\Phi}(X)\widetilde{A}, SB = \widetilde{K}U^T HU\widetilde{K} \in \mathbb{R}^{m\times m}$ and $ST = \widetilde{K}\widetilde{K} \in \mathbb{R}^{m\times m}$. In order to solve the problem (15), we considered two stages: the first stage is to maximize the pseudo between-class scatter matrix SB by QR method and the second stage is to solve a generalized eigenvalue problem. The key problem of the first stage is to deal with the following optimization problem:

$$\hat{A} = \arg\max_{\hat{A}^T\hat{A}=I} \mathrm{tr}(\hat{A}^T SB\hat{A}). \qquad (16)$$

Since H is an $L \times L$ diagonal matrix, it is easy to decompose H into the form $H = H_1 H_1^T$, where $H_1 = \mathrm{diag}(\sqrt{m_1 d_1^{-q}}, \cdots, \sqrt{m_j d_j^{-q}}, \cdots, \sqrt{m_L d_L^{-q}})$, is a $L \times L$ matrix. Consequently, $SB = (\widetilde{K}U^T H_1)(\widetilde{K}U^T H_1)^T = K_1(K_1)^T$, where K_1 is an $m \times L$ matrix.

In general, the number of classes is smaller than the number of training samples. In this case, we can easily prove that $\mathrm{rank(SB)} \leq \mathrm{L} - 1$. When L is much smaller than the number of training samples, we can apply QR technique to decompose K_1 and obtain an efficient method for solving kernel discriminant analysis. In fact, if $K_1 = (Q_1\ Q_2)\binom{R}{0}$ is the QR decomposition of K_1, where $R \in \mathbb{R}^{r\times L}$ is a row full rank matrix, $r = \mathrm{rank}(SB)$ and $Q_1 \in \mathbb{R}^{m\times r}$ and $Q_2 \in \mathbb{R}^{m\times(m-r)}$ are volumn orthogonal matrix, we can verify that Q_1 is a solution of the problem (16).

Theorem 1 *For any orthogonal matrix $G \in R^{r\times r}$, $\hat{A} = Q_1 G$ is a solution of the problem (16).*

Proof: Since $G^T G = GG^T = I_r$ and $Q_1^T Q_1 = I_r$, we have $(Q_1 G)^T(Q_1 G) = I_r$ and

$$\begin{aligned} \mathrm{tr}((Q_1 G)^T SB(Q_1 G)) &= \mathrm{tr}(Q_1^T SBQ_1 GG^T) \\ &= \mathrm{tr}(Q_1^T SBQ_1), \end{aligned}$$

which indicates that the conclusion is true.

Theorem 2 *Let $r = \mathrm{rank}(SB)$ and $K_1 = Q_1 R$ be the QR decomposition of K_1. Let $\widetilde{ST} = Q_1^T STQ_1, \widetilde{SB} = Q_1^T SBQ_1$ and G be a matrix whose columns are the eigenvectors of $(\widetilde{SB})^{-1}\widetilde{ST}$ corresponding to the t largest eigenvalues. Then $Q_1 G$ is an optimal solution of the problem (12).*

Proof: By the QR decomposition of K_1, we know that $\widetilde{SB} = Q_1^T SBQ_1 = R_1 R_1^T$ is nonsingular matrix. According to the definition of the pseudo-inverse of a matrix, we can deduce that

$$\begin{aligned}(SB)^+ &= (K_1(K_1)^T)^+ \\ &= (\left[\begin{array}{cc} Q_1 & Q_2 \end{array}\right] \left[\begin{array}{cc} RR^T & 0 \\ 0 & 0 \end{array}\right] \left[\begin{array}{cc} Q_1 & Q_2 \end{array}\right]^T)^+ \\ &= \left[\begin{array}{cc} Q_1 & Q_2 \end{array}\right] \left[\begin{array}{cc} (RR^T)^{-1} & 0 \\ 0 & 0 \end{array}\right] \left[\begin{array}{cc} Q_1 & Q_2 \end{array}\right]^T\end{aligned}$$

and then

$$\begin{aligned}(SB)^+ STg &= (\left[\begin{array}{cc} Q_1 & Q_2 \end{array}\right] \left[\begin{array}{cc} (RR^T)^{-1} & 0 \\ 0 & 0 \end{array}\right] \\ &\quad \left[\begin{array}{cc} Q_1 & Q_2 \end{array}\right]^T) STg = \lambda g,\end{aligned}$$

which is equivalent to

$$\begin{aligned}&\left[\begin{array}{c} (RR^T)^{-1} \\ 0 \end{array}\right] Q_1^T ST \left[\begin{array}{cc} Q_1 & Q_2 \end{array}\right] \left[\begin{array}{c} Q_1^T \\ Q_2^T \end{array}\right] g \\ &= \lambda \left[\begin{array}{c} Q_1^T \\ Q_2^T \end{array}\right] g.\end{aligned}$$

Hence,

$$(RR^T)^{-1} Q_1^T STQ_1 Q_1^T g = (\widetilde{SB})^{-1} \widetilde{ST} Q_1^T g = \lambda Q_1^T g,$$

which implies that $Q_1^T g$, is a eigenvector of $(\widetilde{SB})^{-1}\widetilde{ST}$ corresponding to the eigenvalue λ. Therefore, the conclusion of the theorem is true. By theorem 3.2, we can obtain the following algorithm.

Algorithm 1 mdkfda/qr algorithm

(1) Compute the fuzzy membership matrix U in terms of minimum distance;

(2) Select the kernel type and compute the kernel matrix K and $\widetilde{K}$;

(3) Calculate matrices $SB = \widetilde{K}U^T HU\widetilde{K} = (\widetilde{K}U^T H_1)(\widetilde{K}U^T H_1)^T = K_1(K_1)^T$ and $ST = \widetilde{K}\widetilde{K}$;

(4) Compute the QR decomposition of K_1: $K_1 = Q_1 R$, let $\widetilde{ST} = Q_1^T STQ_1$ and $\widetilde{SB} = Q_1^T SBQ_1$;

(5) Compute the eigenvectors G, of the matrix $(\widetilde{SB})^{-1}\widetilde{ST}$ corresponding to the t largest eigenvalues, let$\widetilde{A} = Q_1 G$;

(6) For any input vector x, its low dimensional feature representation by mdkfda/qr is

$$\begin{aligned}y_x &= \widetilde{A}^T \widetilde{\Phi}(X)^T \phi(x) \\ &= G^T Q_1^T (I - 1_{m\times m}/m)^T \\ &\quad (k(x_1, x), k(x_2, x) \cdots, k(x_m, x))^T.\end{aligned}$$

4 Experiments and analysis

we evaluate the performance of mdkfda/qr algorithm on face recognition task. The publicly available face databases, namely ORL database and a subset of the FERET database are used for experiments.

All experiments are performed on a PC (2.40 GHZ CPU, 2G RAM) with MATLAB 7.1. Because training sets are obtained randomly in experiments. Three discriminant analysis-based feature extraction methods, namely the proposed mdkfda/qr, KDA [15] and FDA [27] are tested and compared. For each of the three methods, the face recognition procedure consists of: (i) a feature extraction step where three kinds of feature representation of each training or test sample are extracted by mdkfda/qr, KDA and FDA, respectively; (ii) The nearest neighbor classifier is used.

It is known that proper selection of kernel function is important to achieve better performance in kernel-based learning methods. Generally speaking, there are two classes of widely used kernel functions: polynomial kernel and Gaussian kernel. In order to evaluate the effect and stable QR decomposition in mdkfda/qr algorithm, we take into consideration polynomial kernel (17) and Gaussian kernel (18).

$$k(x, y) = (x \cdot y + 1)^p. \tag{17}$$

$$k(x, y) = exp(-\|x - y\|^2/2\sigma^2). \tag{18}$$

The parameter p is set as $2, \cdots, 6$, and the parameter σ is set as $6, \cdots, 12$, respectively. We then tested the proposed mdkfda/qr, KDA and FDA with different parameters p, σ, and q of weighting function.

4.1. Experiment on the ORL database

The ORL database contains 40 persons, each having 10 different images. The images of the same person are taken at different times under slightly varying lighting conditions and with various facial experiments. Some people are captured with or without glasses. The heads in images are slightly titled or rotated. The images in the database are manually cropped and recalled to 112×92. In order to reduce the size of the image, we obtain the size of 28×23 pixels. So, the number of features of each character is 644. In the experiments, 8 images are randomly taken from 10 images as training samples sets and the rest are used testing sets. In order to make full use of the available data and to evaluate the generalization power of the algorithms more accurately, we adopt across-validation strategy and run the system 30 times. Figure 1 shows several sample images of some persons in ORL. The weighting function $w(d_0) = (D)^{-q}$ was used in the experiments. In ORL dataset, parameter $q = 2$, and polynomial kernel were employed, the result is shown in Figure 2. From Fig.2, we clear to see that the proposed mdkfda/qr approach outperforms KDA and FDA in terms of the accuracy of classification. In addition, mdkfda/qr and KDA methods are insensitive to the parameter p of the polynomial

Fig.1. Sample images of some persons in the ORL database.

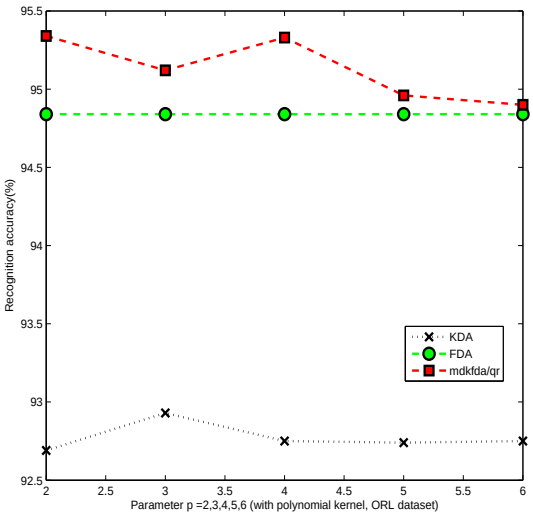

Fig.2. Mean correct recognition rate curves with the polynomial kernel on ORL database.

kernel function, meanwhile, mdkfda/qr method can achieve the best recognition accuracy with the parameter $p = 2$. However, for FDA method, there is no change according to the accuracy of classification.

In order to improve the performance of classification and evaluate the effective of QR decomposition, the Gaussian kernel with the parameters σ is set as $6, \cdots, 12$ and the $q = 2$ of weighted function (Viz. $w(d_0) = (D)^{-2}$ are used in our experiments. The experiment results are shown in Fig.3. According to Fig.3, we found that the proposed mdkfda/qr approach outperforms KDA and FDA in terms of the accuracy of classification. In addition, for KDA method, as the parameter σ rises, the recognition accuracy is declining. But, there is no change in FDA approach according to accuracy of classification. That is because there is no parameter was used in FDA method. Meanwhile, mdkfda/qr approach gives best result with $\sigma = 6$ in terms of recognition accuracy.

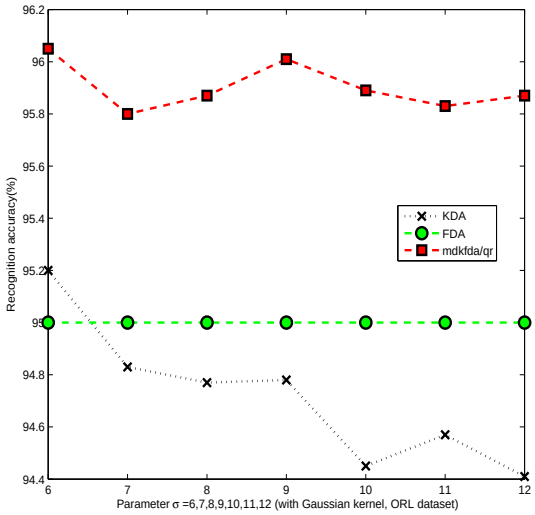

Fig.3. Mean correct recognition rate curves with the Gaussian kernel on ORL database.

In the following experiment, we will explore the influence of different weighted functions. In order to evaluate the weighted function $w(d_0) = (D)^{-q}, (q \geq 2)$ also influence the accuracy of classification in mdkfda/qr approach, we further compare the mdkfda/qr method on the ORL database. The 2-polynomial kernel (Viz. $p = 2$) and 6-Gaussian kernel (Viz. $\sigma = 6$) were used in the experiments. The results are shown in Table 1.

Table 1: Mean correct recognition rate (%) of mdkfda/qr method on ORL (polynomial and Gaussian kernels).

Parameter	2	4	6	8
$p = 2$	95.34	94.87	95.68	94.96
$\sigma = 6$	95.75	95.78	95.86	95.85

From Table 1, we can see that mdkfda/qr method gives better results with Gaussian kernel than with polynomial kernel in terms of correct recognition rate.

4.2.Experiment on a subset of FERET database

The FERET face database is a result of the FERET program, which was sponsored by the US Department of Defense through the DARPA Program. It has become a standard database for testing and evaluating state of the art face recognition algorithms. The proposed method was tested on a subset of the FERET database. This subset includes 1000 images of 200 in-

dividuals (each individual has five images). This subset involves variations in facial expression, illumination, and pose. In our experiment, the facial portion of each original images was automatically cropped images was resized to 20×20 pixels. We split the whole database into two parts evenly, 2 images are randomly taken from 5 images as training samples sets and the rest are used testing sets. In order to make full use of the available data and to evaluate the generalization power of algorithms more accurately, we adopt a across-validation strategy and run the system thirty times. Figure 4 shows several sample images of some persons in FERET. The weighting function $w(d_0) = (D)^{-q}$ was used in the experiments. In FERET dataset, parameter $q = 2$, and polynomial kernel were employed, the result is shown in Figure 5.

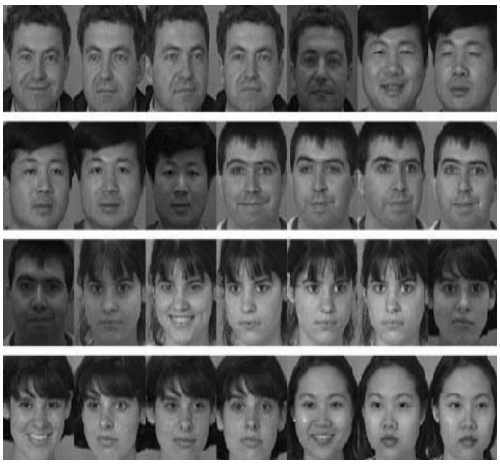

Fig.4. Sample images of some persons in FERET dataset.

According to Fig.5, we can see that the proposed mdkfda/qr approach outperforms KDA and FDA in terms of the accuracy of classification. In addition, mdkfda/qr method is insensitive to the parameter p of the polynomial kernel function, meanwhile, mdkfda/qr method can achieve the best recognition accuracy with the parameter $p = 5$. However, for FDA method, there is no change according to the accuracy of classification (Because there is no parameter was used in FDA method). There is a huge difference between KDA and mdkfda/qr and FDA in terms of accuracy of classification.

Similarly, in order to improve the performance of classification, the Gaussian kernel with the parameters σ is set as $6, \cdots, 12$ and the $q = 2$ of weighted function (Viz. $w(d_0) = (D)^{-2}$ are used in our experiments, and the results are shown in Fig.6.

Fig.6 shows that the proposed mdkfda/qr approach outperforms KDA and FDA in terms of the accuracy of classification. In addition, for mdkfda/qr, KDA and FDA methods are not insensitive to the parameter σ. There is a huge difference between KDA and mdkfda/qr and FDA in terms of accuracy of classification.

In the following experiment, the 2-polynomial kernel (Viz. $p = 2$) and 6-Gaussian kernel (Viz. $\sigma = 6$) were used in the experiments. The results are shown in Table 2.

From Table 2, we can see that mdkfda/qr method gives better results with Gaussian kernel than with

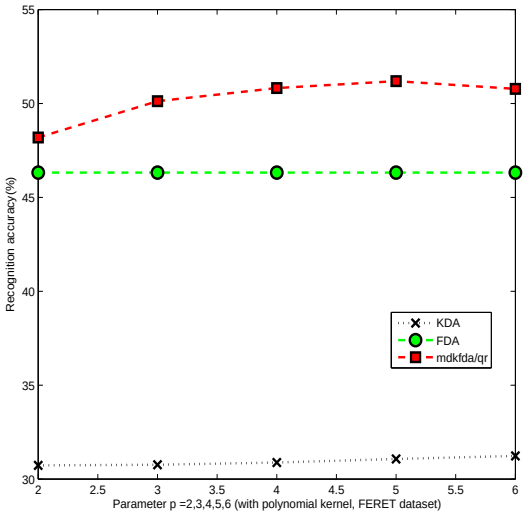

Fig.5. Mean correct recognition rate curves with the polynomial kernel on FERET database.

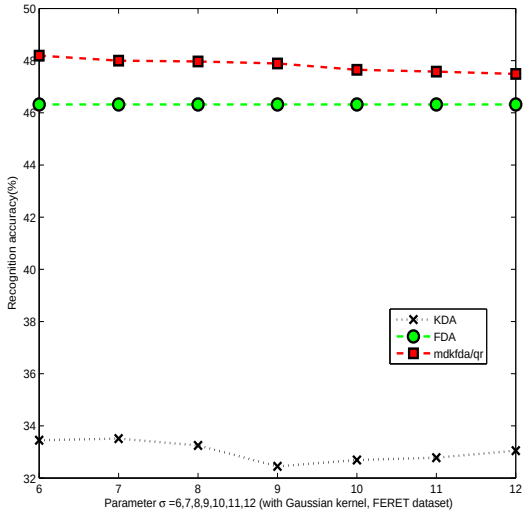

Fig.6. Mean correct recognition rate curves with the Gaussian kernel on FERET database.

Table 2: Mean correct recognition rate (%) of mdkfda/qr method on FERET (polynomial and Gaussian kernels).

Parameter	2	4	6	8
$p = 2$	48.19	50.15	52.20	51.20
$\sigma = 6$	50.08	50.15	51.25	52.30

polynomial kernel in terms of correct recognition rate.

5 Conclusion

In this paper, we present a kernel fuzzy discriminant analysis minimum distance-based approach for the classification of face images. This approach can find lower dimensional nonlinear features with significant discriminant power and can be viewed as a kernel generalization of FDA. Experiments show that QR decomposition is an efficient and effective step and then mdkfda/qr algorithm is effective and feasible in real world application. In order to compare mdkfda/qr, FDA and KDA methods, we select different parameters of kernel functions and weighted function. Experimental result shows that the selection of parameters is very important in terms of accuracy of classification. For the ORL database and FERET database, Gaussian kernel function is the best according to classification rate. So, mdkfda/qr is more feasible than FDA and KDA.

The future works on this subject will have to investigate the influence of parameter, distance and kernel function in the face recognition problems.

Acknowledgements: The authors are very grateful to the editor and anonymous referees reviews for their valuable comments and helpful suggestions. In addition, this work is supported by the Graduates' Research Innovation Program of Higher Education of Jiangsu Province (Grant No. CXZZ13-0239). The corresponding author of this paper is: Jianqiang Gao.

References:

[1] Martnez AM, Kak AC, PCA versus LDA, *IEEE Trans Pattern Anal Mach Intell.* 23 (2001), pp. 228–233.

[2] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, A. Smola, Muller KR Constructing descriptive and discriminative nonlinear features: Rayleigh coefficients in kernel feature spaces, *IEEE Trans Pattern Anal Mach Intell.*, 25 (2003), pp. 623–628.

[3] T. Hastie, R. Tibshirani, J.H. Friedman, The elements of statistical learning: data mining, inference, and prediction, *Springer*. (2001).

[4] M. Loog, R.P.W. Duin, R. Hacb-Umbach, Multiclass linear dimensuin reduction by weighted pairwise fisher criteria, *pattern recognition and machine intelligence*, 23 (2001), pp. 762–766.

[5] H.M. Lec, C.M. Chen, 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), pp. 426–432.

[6] D.L. Swets, J. Weng, Using discriminant eigenfeatures for image retrieval, *IEEE Trans Pattern Anal Mach Intell.*, 18(8), (1996), pp. 831–836.

[7] V. Belhumeur, J. Hespanha, D. Kriegman, Eigenfaces vs fisherfaces: recognition using class specific linear projection, *IEEE Trans Pattern Anal Mach Intell.*, 19(7), (1997), pp. 711–720.

[8] J. Yang, J.Y. Yang, Why can LDA be performed in PCA transformed space? *Pattern Recog.*, 36(2), (2003), pp. 563–566.

[9] J. Ye, Q. Li, LDA/QR: An efficient and effective dimension reduction algorithm and its theoretical foundation, *Pattern Recog.*, 37 (2004), pp.851–854.

[10] J. Gao, L. Fan, The impact on the classification results from distances in weighted PCA and LDA (in chinese), *Journal of Liaocheng University (Natural science)*, 23(4), (2010), pp. 4–8.

[11] J. Gao, Liya Fan, The impact on the face recognition from distances in fuzzy linear discriminant analysis, (in chinese), *Journal of Jinggangshan University (Natural Science)*, 33(3), (2012), pp. 1–7.

[12] J. Gao, L. Fan, L. Xu, Solving the face recognition problem using QR factorization, *WSEAS Transactions on Mathematics*, 8(11), (2012), pp. 728–737.

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

[14] V.N. Vapnik, Statistical learning theory, *wiley, New York*, (1998).

[15] B. Scholkopf, A.J. Smola, K.-R. Muller, Nonlinear component analysis as a kernel eigenvalue problem, *Neural Comput.*, 10 (1998), pp. 1299–1319.

[16] Y.Y. Liu, X.P. Liu, Z.X. Su, A new fuzzy approach for handing class labels in canonical correlation analysis, *Neural Comput.*, 71 (2008), pp. 1735–1740.

[17] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, Muller KR Fisher Discriminant Analysis with Kernels. *In: Proceedings of IEEE international workshop neural networks for signal processing IX*, (1999), pp. 41–48.

[18] J. Yang., A.F. Frangi, J.Y. Yang, D. Zhang, KPCA plus LDA: a complete kernel Fisher discriminant framework for feature extraction and recognition, *IEEE Trans Pattern Anal Mach Intell.*, 27 (2005), pp. 230–244.

[19] G. Dai, Y.T. Qian, S. Jia, A kernel fractional-step nonlinear discriminant analysis for pattern recognition. *In: Proceedings of the 18th international conference on pattern recognition*, (2004), pp. 431–434.

[20] G. Dai, D.Y. Yeung, Y.T. Qian, Face recognition using a kernel fractional-step discriminant analysis algorithm, *Pattern Recogn.*, 40 (2007), pp. 22–243.

[21] D. Zhou, Z. Tang, Kernel-based improve discriminant analysis and its applications to faces recognition, *soft comput.*, 14 (2010), pp.102–111.

[22] D. Zhou, Z. Tang, A modification of kernel discriminant analysis for high-dimensional data-with application to face recognition, *signal processing*, 90 (2010), pp. 2423–2430.

[23] R. Lotlikar, R. Kothari, Fractional-step dimension reduction, *IEEE Trans Pattern Anal Mach Intell.*, 22 (2000), pp. 623–627.

[24] L.A. Zadeh, Fuzzy sets, *Information Control*, 8 (1965), pp. 338–353.

[25] K.C. Kwak, W.Pedrycz, Face recognition using a fuzzy Fisher-face classifier, *Pattern Recog.*, 38(10), (2005), pp. 1717–1732.

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