Pattern Recognition 46 (2013) 2599–2611

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Evaluation of weighted Fisher criteria for large category dimensionality reduction in application to Chinese handwriting recognition

Xu-Yao Zhang *, Cheng-Lin Liu

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, No. 95 Zhongguancun East Road, Beijing 100190, P.R. China

ARTICLE INFO

Article history:
Received 7 July 2012
Received in revised form
12 December 2012
Accepted 30 January 2013
Available online 24 February 2013

Keywords:
Dimensionality reduction
Large category
Class separation problem
Weighted Fisher criteria
Class level
Sample level
Chinese handwriting recognition

ABSTRACT

To improve the class separability of Fisher linear discriminant analysis (FDA) for large category problems, we investigate the weighted Fisher criterion (WFC) by integrating weighting functions for dimensionality reduction. The objective of WFC is to maximize the sum of weighted distances of all class pairs. By setting larger weights for the most confusable classes, WFC can improve the class separation while the solution remains an eigen-decomposition problem. We evaluate five weighting functions in three different weighting spaces in a typical large category problem of handwritten Chinese character recognition. The weighting functions include four based on existing methods, namely, FDA, approximate pairwise accuracy criterion (aPAC), power function (POW), confused distance maximization (CDM), and a new one based on K-nearest neighbors (KNN). All the weighting functions can be calculated in the original feature space, low-dimensional space, or fractional space. Our experiments on a 3,755-class Chinese handwriting database demonstrate that WFC can improve the classification accuracy significantly compared to FDA. Among the weighting functions, the KNN method in the original space is the most competitive model which achieves significantly higher classification accuracy and has a low computational complexity. To further improve the performance, we propose a nonparametric extension of the KNN method from the class level to the sample level. The sample level KNN (SKNN) method is shown to outperform significantly other methods in Chinese handwriting recognition such as the locally linear discriminant analysis (LLDA), neighbor class linear discriminant analysis (NCLDA), and heteroscedastic linear discriminant analysis (HLDA).

1. Introduction

In pattern classification for high-dimensional data, it is common to apply a feature extraction method as a pre-processing technique, not only to reduce the computational complexity, but also to obtain better generalization performance, by reducing irrelevant and redundant information in the data, and overcoming the estimation problem in statistical classifier learning. The feature extraction methods include linear and nonlinear dimensionality reduction ones. A large variety of linear methods, such as principal component analysis (PCA) [19,41], Fisher linear discriminant analysis (FDA) [12], independent component analysis (ICA) [18], non-negative matrix factorization (NMF) [24] and locality preserving projections (LPP) [16], have been proposed from different statistical or geometrical viewpoints. The nonlinear methods include (i) the kernel extension of the linear methods, such as kernel PCA [36] and kernel FDA [45]; (ii) manifold learning models such as the ISOMAP [40], LLE [35] and Laplacian eigenmaps [3]; (iii) deep neural networks [17,43,38] which use a deep architecture to learn the nonlinear data mapping. The dimensionality reduction methods can also be divided into data-independent (e.g. random projection [9,6]), unsupervised (e.g. PCA, NMF, LPP), and supervised (e.g. FDA, kernel FDA), according to the increasing level of involvement of the data.

In this paper, we focus on supervised linear dimensionality reduction for large category problems, which incur high computational complexity to nonlinear methods and even some linear methods. The most well-known supervised linear method is the FDA, which was first developed by Fisher [11] for binary classification and then extended by Rao [34] to multi-class problems. The purpose of linear dimensionality reduction is to learn a transformation matrix $W \in \mathbb{R}^{d \times d'}$ to transform the feature from $\mathbb{R}^d$ into a low-dimensional space $\mathbb{R}^{d'}$ ($d' < d$). The objective of FDA is to maximize the between-class distance as well as minimize the within-class distance. FDA is the optimal model for linear dimensionality reduction [12], when (i) the class-conditional distribution is Gaussian with equal covariance matrix for all the classes (homoscedastic); and (ii) the reduced dimensionality is $C-1$ (C is the number of classes).

For large category problems where $C \gg d > d'$, however, FDA suffers from the class separation problem. The objective of FDA can be formulated as maximizing the sum of all the pairwise

* Corresponding author. Tel.: +86 15011322387; fax: +86 10 62551993.
E-mail addresses: xyz@nlpr.ia.ac.cn (X.-Y. Zhang), liucl@nlpr.ia.ac.cn (C.-L. Liu).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.01.036

distances between different classes, which overemphasizes the large distances of the already well-separated classes, and confuses the classes that are close in the original feature space. Many methods have been proposed to overcome this problem. Loog et al. [32] proposed the approximate pairwise accuracy criterion (aPAC), which uses a weighting function to emphasize the close class pairs in the between-class scatter matrix. Lotlikar and Kothari [33] developed the fractional-step FDA, which is also a weighting approach but selects a subspace through fractional steps. Instead of the arithmetic mean of distances used in FDA, Tao et al. [39] proposed to use the geometric mean, while Bian and Tao [4] proposed to use the harmonic mean, which require complex computation in subspace solution, however. Recently, the idea of maximizing the minimal pairwise distance was proposed to solve the class separation problem [50,44,47,5]. Simultaneously maximizing all the pairwise distances was also proposed as a multi-objective optimization problem [2] to handle the class separation problem. Besides the class separation problem, many other methods have been developed to deal with the heteroscedastic problem [31,51], to extract more than $C-1$ features for small category problems [21], and to alleviate the small sample size problems [7,46].

Handwritten Chinese character recognition is a typical large category problem where FDA has been popularly used for dimensionality reduction (e.g., [14,25]). The linear dimensionality reduction methods based on weighted Fisher criterion (WFC), such as the aPAC and the confused distance maximization (CDM) [48], are applicable to large category problems with moderate computation cost because the subspace solution remains an eigen-decomposition problem. To improve the class separability of FDA for large category problems, we evaluate various weighting functions in different feature spaces. The weighting functions include four based on existing methods, namely, the FDA (a special case of WFC), the aPAC, the power function (POW) used in fractional-step dimensionality reduction [33], and the CDM. We explore a new model which maximizes the sum of the distances between each class and its k-nearest neighbors (KNN). The weighting functions are calculated in three different spaces: the original feature space, the low-dimensional space, and the fractional space, which have increasing computational complexities but lead to better approximations of the weighting function in the final reduced space. We compare the five weighting functions from four perspectives: (i) the class separation in the reduced subspace; (ii) the locality; (iii) the property of classifier-dependence and (iv) the property of space invariance. We evaluated the weighting functions and weighting spaces on a 3,755-class Chinese handwriting dataset. The experimental results show significant improvement of classification accuracy of WFC over the ordinary FDA. The KNN weighting function in the original space is the most competitive model in respect of both the classification accuracy and the computational complexity. To the best of our knowledge, this is the first work on the evaluation of different weighting functions in different weighting spaces of WFC for large category dimensionality reduction.

Another contribution of this paper is the extension of the KNN based weighted Fisher criterion from class level to sample level (denoted as sample-level KNN: SKNN). SKNN is a nonparametric extension of the KNN method. By computing the between-class scatter matrix at sample level, SKNN can capture much more information of the decision boundary, solve the class separation problem, and also alleviate the heteroscedastic and multi-modal problems. Compared with some popular methods in Chinese handwriting recognition such as locally linear discriminant analysis (LLDA) [15], neighbor class linear discriminant analysis (NCLDA) [42], and heteroscedastic linear discriminant analysis (HLDA) [30,31], SKNN can achieve much higher classification accuracy for both the nearest class mean (NCM) and modified quadratic discriminant function (MQDF) [20] classifiers on all the reduced subspaces consistently. All the compared results can be exactly repeated with the feature data released at [1], and therefore can be used as a benchmark for comparing different dimensionality reduction models for large category problems.

The rest of this paper is organized as following: Section 2 introduces the class separation problem of FDA and some related works attempting to solve this problem; Section 3 presents the framework of weighted Fisher criteria (WFC), and describes five weighting functions in three different weighting spaces; Section 4 reports comprehensive evaluations of different WFC from the aspects of accuracy, complexity, statistical significance, space invariance, and similar characters; Section 5 extends the KNN method to the sample level, and compares the new method with some popular methods in Chinese handwriting recognition; and Section 6 draws concluding remarks.

2. FDA and class separation problem

Let $\mu_i \in \mathbb{R}^d$ and $\Sigma_i \in \mathbb{R}^{d\times d}$ be the mean vector and the covariance matrix for class i, $(i = 1 \cdots C)$, respectively. The within-class and between-class scatter matrices are defined as:

$$S_w = \sum_{i=1}^{C} p_i \Sigma_i, \tag{1}$$

$$S_b = \sum_{i,j=1}^{C} p_i p_j (\mu_i - \mu_j)(\mu_i - \mu_j)^\top, \tag{2}$$

where $p_i = N_i/N$, $N = \sum_{i=1}^{C} N_i$ (N_i is the number of samples in class i). The objective of FDA is to learn a transformation matrix $W \in \mathbb{R}^{d\times d'}$ $(d' < d)$ to transform the feature vector $x \in \mathbb{R}^d$ into a low-dimensional vector $x' \in \mathbb{R}^{d'}$ as $x' = W^\top x$ by minimizing the within-class variance while maximizing the between-class variance. It is easy to verify that the scatter matrices in the transformed space become $W^\top S_w W$ and $W^\top S_b W$. There are many formulations of FDA, and two typical criteria are given below [12]:

$$\max_W \mathrm{tr}\{(W^\top S_w W)^{-1}(W^\top S_b W)\}, \tag{3}$$

$$\max_W \{\ln|W^\top S_b W| - \ln|W^\top S_w W|\}, \tag{4}$$

which are equivalent to a constrained problem:

$$\max_{W_{\mathrm{FDA}} \in \mathbb{R}^{d\times d'}} \mathrm{tr}(W_{\mathrm{FDA}}^\top S_b W_{\mathrm{FDA}}) \quad \text{s.t. } W_{\mathrm{FDA}}^\top S_w W_{\mathrm{FDA}} = I, \tag{5}$$

where I is the identity matrix. Usually, this model is solved by an equivalent two-step approach: whitening followed by PCA in the whitened space.

2.1. The first step: whitening

Let P be the eigenvector matrix and Λ be the diagonal eigenvalue matrix of the within-class scatter matrix:

$$S_w = P\Lambda P^\top. \tag{6}$$

The whitening transformation is defined as:

$$W_{\mathrm{whiten}} = P\Lambda^{-1/2} \in \mathbb{R}^{d\times d}, \tag{7}$$

which satisfies

$$W_{\mathrm{whiten}}^\top S_w W_{\mathrm{whiten}} = I. \tag{8}$$

It is implicitly assumed that within-class scatter matrix S_w is invertible. For the large category problem with enough training samples, this assumption can be generally guaranteed. In the case

of singular S_w, the zero values in Λ can be set as some small non-zero positive constant.

2.2. The second step: PCA in the whitened space

Defining the transformation matrix of FDA as

$$W_{\text{FDA}} = W_{\text{whiten}} W, \tag{9}$$

and inserting this equation into the objective function of FDA (5), the problem becomes

$$\max_{W \in \mathbb{R}^{d\times d'}} \text{tr}(W^\top W_{\text{whiten}}^\top S_b W_{\text{whiten}} W) \quad \text{s.t. } W^\top W = I. \tag{10}$$

This problem is equivalent to

$$\max_{W \in \mathbb{R}^{d\times d'}} \sum_{i,j=1}^{C} p_i p_j \triangle_{ij} \quad \text{s.t. } W^\top W = I, \tag{11}$$

where $\triangle_{ij}$ is the distance between the class means of classes i and j in the transformed subspace

$$\triangle_{ij} = \|W^\top W_{\text{whiten}}^\top (\mu_i - \mu_j)\|_2^2. \tag{12}$$

The second step of FDA is thus to solve (11), which is exactly the principal component analysis (PCA) among whitened class means $W_{\text{whiten}}^\top \mu_1, \ldots, W_{\text{whiten}}^\top \mu_C$ (Fig. 1(b)). However, PCA is a global model, which maximizes the sum of all the pairwise distances. The local information for distinguishing one class from another may be lost after PCA transformation, which results in the class separation problem.

2.3. The class separation problem

The first step of FDA is to learn a suitable distance metric: in the whitened space, the Euclidean distance becomes the optimal measurement. In the second step, because (11) is to maximize the sum of all the pairwise distances, it will cause the class separation problem [32]. To illustrate this, consider that one class is located remotely from the other classes and can be considered as an outlier (Fig. 1(c)). In this case, by optimizing (11), the projection axis of FDA is the one that separates the outlier from the remaining classes as much as possible. The pairs of large-distance classes completely dominate the solution of (11). As a consequence, there is a large overlap among the remaining classes, leading to an overall low and suboptimal classification performance.

To solve the class separation problem, Tao et al. [39] proposed to maximize the geometric mean $\{\max \sum_{i\neq j} p_i p_j \log \triangle_{ij}\}$. Bian and Tao [4] further proposed to maximize the harmonic mean $\{\max -\sum_{i\neq j} p_i p_j \triangle_{ij}^{-1}\}$. Recently, many authors have proposed to maximize the minimal distance $\{\max(\min_{i\neq j} \triangle_{ij})\}$ [50,44,47,5]. Abou-Moustafa et al. [2] further proposed to maximize all the pairwise distances $\{\max \triangle_{12}, \max \triangle_{13}, \ldots, \max \triangle_{C-1,C}\}$ simultaneously in multi-objective optimization. Although these methods have reported improved performance, they are all based on some complex iterative optimization procedures (Table 1), which make them not scalable for large category (e.g. thousands of classes) problems.

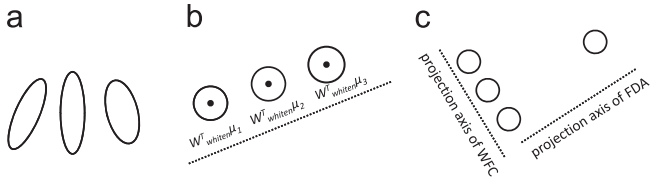


Fig. 1. (a) The distributions of three classes; (b) After whitening transformation, each class can be approximately represented by a sphere; (c) An illustration of the class separation problem of FDA.

Table 1
Optimization methods and experimental datasets used by different models.

Method	Optimization	Experiments (C classes)
[39]	steepest gradient	UCI and USPS ($C \le 10$)
[4]	conjugate gradient	UCI and Objects ($C \le 20$)
[50]	constrained concave-convex procedure (CCCP)	UCI and Face ($C \le 100$)
[44]	semi-definite programming (SDP)	UCI and Face ($C \le 40$)
[5]	sequential SDP	UCI and Face ($C \le 50$)
[2]	gradient descent	Image and UCI ($C \le 40$)

Another widely used method to solve the class separation problem is the weighted Fisher criterion (WFC) $\{\max \sum_{i,j=1}^{C} f_{ij} p_i p_j \triangle_{ij}\}$ [32,33,48]. By integrating the weighting function f_{ij} into the objective function of (11) and setting larger weights for the most confusable classes, WFC can solve the class separation problem effectively. Furthermore, the solution of WFC only requires solving an eigen-decomposition problem and no complex iterative optimization is needed, which makes WFC efficient for large scale applications. In this paper, we investigate the WFC with various weighting functions for large category classification. Specifically, we evaluate the classification accuracies and running times of five weighting functions in three different weighting spaces on a large scale 3,755-class handwriting dataset.

3. Weighted Fisher criteria

To solve the class separation problem, the objective function of (11) is generalized by introducing a weighting function, resulting in the weighted Fisher criterion (WFC):

$$\max_{W \in \mathbb{R}^{d\times d'}} \sum_{i,j=1}^{C} f_{ij} p_i p_j \triangle_{ij} \quad \text{s.t. } W^\top W = I, \tag{13}$$

where $f_{ij} \ge 0$ is a weighting function that depends on the probability of confusion or merging in the reduced subspace between class i and class j. By setting larger f_{ij} for the class pairs which are closer together and likely to cause confusion, WFC helps make the optimality criteria more representative of the classification ability in the reduced space.

The criterion of (13) can be re-written as:

$$\max_{W \in \mathbb{R}^{d\times d'}} \text{tr}(W^\top \widehat{S}_b W) \quad \text{s.t. } W^\top W = I, \tag{14}$$

where $\widehat{S}_b$ is the weighted between-class scatter matrix in the whitened space:

$$\widehat{S}_b = \sum_{i,j=1}^{C} f_{ij} p_i p_j (\widehat{\mu}_i - \widehat{\mu}_j)(\widehat{\mu}_i - \widehat{\mu}_j)^\top, \tag{15}$$

and $\widehat{\mu}_i$ is the whitened mean for class i:

$$\widehat{\mu}_i = W_{\text{whiten}}^\top \mu_i \ \forall i = 1,2,\ldots,C. \tag{16}$$

The model of (14) can be solved by taking the columns of the $d \times d'$ matrix W to be the d' eigenvectors corresponding to the d' largest eigenvalues of $\widehat{S}_b$. Denote the solution of WFC in (14) as $W_{\text{WFC}} \in \mathbb{R}^{d\times d'}$, the final dimensionality reduction matrix is the accumulation of the whitening transformation (7) and WFC:

$$W_{\text{final}} = W_{\text{whiten}} W_{\text{WFC}} \in \mathbb{R}^{d\times d'}. \tag{17}$$

The weighted Fisher criterion (WFC) is a general framework. Its performance depends on the definition of the weighting matrix

$$F = \{f_{ij}\} \in \mathbb{R}^{C\times C}. \tag{18}$$

3.1. Weighting function

In this section, we describe and compare five weighting functions, which can be used in the WFC framework for dimensionality reduction.

3.1.1. FDA

Clearly, the ordinary FDA is a special case of WFC adopting a constant weighting function

$$\text{FDA}: f_{ij}=1, \forall i,j=1,\ldots,C. \tag{19}$$

Since all the class pairs are equally weighted, FDA will overemphasize the large-distance class pairs and cause an overlapping of the small-distance class pairs.

3.1.2. aPAC

Loog et al. [32] proposed the approximate pairwise accuracy criterion (aPAC) by using a weighting function as

$$\text{aPAC}: f_{ij}=\frac{1}{2d_{ij}^2}\,\text{erf}\left(\frac{d_{ij}}{2\sqrt{2}}\right), \tag{20}$$

where $\text{erf}(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,dt\in[-1,1]$ is the error function[1], and d_{ij} is the distance between class i and class j in the whitened space:

$$d_{ij}=\|\widehat{\mu}_i-\widehat{\mu}_j\|_2=\|W_{\text{whiten}}^{\top}(\mu_i-\mu_j)\|_2. \tag{21}$$

The aPAC is derived to approximate the Bayes error for class pairs, and can solve the class separation problem by setting larger f_{ij} for small-distance (small d_{ij}) class pairs.

3.1.3. POW

Lotlikar and Kothari [33] proposed the fractional-step dimensionality reduction model which uses a weighting function as:

$$\text{POW}: f_{ij}=d_{ij}^{-m}, \tag{22}$$

where m is a positive integer. The dropping of f_{ij} should be faster than the increasing of d_{ij}, and so m is suggested to be $m\geq 3$. Since Eq. (22) is a power function, we denote this method by POW.

3.1.4. CDM

Zhang and Liu [48] proposed the confused distance maximization (CDM) to solve the class separation problem, which defines the weighting function as the confusion probability among different classes:

$$\text{CDM}: f_{ij}=\begin{cases}\dfrac{N_{i\to j}}{N_i}, & i\neq j\\ 0, & i=j\end{cases} \tag{23}$$

where N_i is the number of samples in class i, and $N_{i\to j}$ is the number of samples of class i that are classified into class j by a specific classifier. To obtain a better generalization performance, the confusion matrix $F=\{f_{ij}\}\in\mathbb{R}^{C\times C}$ should be estimated from a dataset which is different from the one used to train the basic classifier, e.g., using the holdout-validation or cross-validation. The weighting function of CDM is defined as the confusion probability estimated from the data with a pre-learned classifier, and therefore is more relevant to the classification task. CDM has shown better classification performance than FDA, aPAC and POW [48].

3.1.5. KNN

A new weighting function named KNN is proposed in this paper to maximize the sum of the distances between each class and its k nearest neighbors:

$$\text{KNN}: f_{ij}=\begin{cases}1, & \text{if } \widehat{\mu}_j\in\text{KNN}(\widehat{\mu}_i)\\ 0, & \text{otherwise}\end{cases} \tag{24}$$

where $\text{KNN}(\widehat{\mu}_i)$ denotes the k nearest neighbors of the whitened class mean $\widehat{\mu}_i$ in $\{\widehat{\mu_1},\ldots,\widehat{\mu_{i-1}},\widehat{\mu_{i+1}},\ldots,\widehat{\mu_C}\}$. The benefits of considering each class with its k nearest neighbors include: (i) focusing on the nearest class pairs, and removing the influence of the large-distance class pairs; (ii) the geometric relationship of different classes is preserved by the connection and propagation between each class and its nearest neighbors; (iii) the fast construction and sparsity of the weighting matrix can significantly reduce the computational complexity of WFC (Section 4.7); (iv) the KNN weighting matrix is nearly space invariant, which means the KNN relationship between the class pairs is nearly the same either in the original feature space or the low-dimensional subspace (Section 4.9).

3.1.6. Comparison of weighting functions

We show different weighting functions in Fig. 2 and make qualitative comparisons from the following perspectives (Table 2):

- **Class separation.** By setting larger weights for the most confusable classes, the weighting functions of aPAC, POW, CDM and KNN have the ability to solve the class separation problem caused by the constant weighting function in FDA.
- **Locality.** The weighting matrices of aPAC and POW are based on the pairwise distance d_{ij}. CDM is based on the confusion matrix, and since each class is only confused with a small number of classes, the weighting matrix of CDM is very sparse. Furthermore, for the KNN weighting matrix, there are at most k non-zero elements for each row (each class). Therefore, the weighting matrices of CDM and KNN are much more sparse (local) than aPAC and POW. The locality property makes CDM and KNN more relevant to the classification accuracy by focusing on the most confusable classes. Furthermore, the computation cost of $\widehat{S}_b$ in Eq. (15) will be significantly reduced due to the locality of f_{ij}.
- **Classifier-dependence.** Fig. 2 shows the CDM weighting matrices for two different classifiers (nearest class mean (NCM) and modified quadratic discriminant function (MQDF)) described in Section 4.2. We can see that the confusion matrix used in CDM is classifier-dependent, which indicates that CDM may be more relevant to a particular classifier, while the weighting matrices of FDA, aPAC, POW and KNN are independent of the classifier.
- **Space invariance.** Another important property of the weighting function is space invariance, which means the weighting functions in the original feature space and in the reduced subspace should be close to each other. The weighting function of FDA is space invariant, and the one of KNN is nearly space invariant due to the geometry preserving of K-nearest neighbors. This will be shown in detail in Section 4.9, and in the following section we will show that this property is very important for dimensionality reduction.

3.2. Weighting space

The weighting function can be defined in different spaces. Because we want to learn a transformation from $\mathbb{R}^d$ to $\mathbb{R}^{d'}$ and the classification performance is directly evaluated in $\mathbb{R}^{d'}$, the optimal weighting function should be defined in the **final reduced space** (FRS) $\mathbb{R}^{d'}$ to reflect the real confusion relationship between classes. However, the weighting function definition in FRS and the transformation matrix learning of WFC are two problems with

[1] http://wikipedia.org/wiki/Error_function

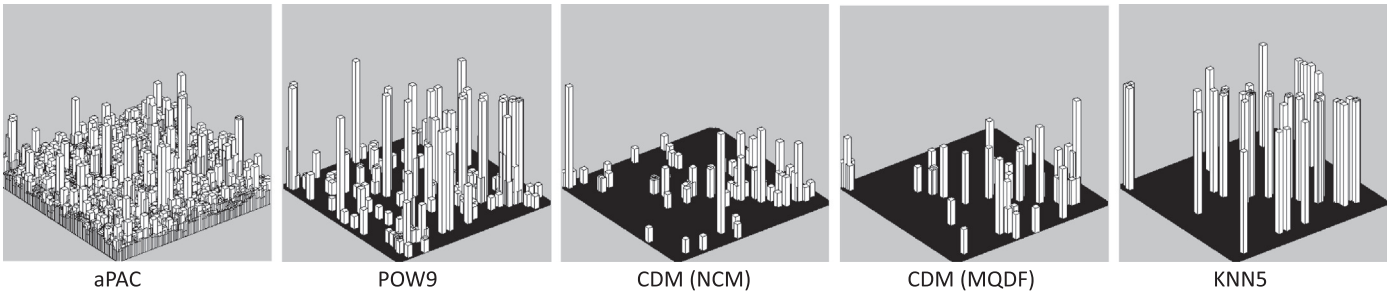


Fig. 2. The 100×100 weighting matrix of different methods for the first 100 classes of the 3,755-class problem.

Table 2
Comparison of different weighting functions.

	class separation	locality	classifier-dependence	space invariance
FDA				√
aPAC	√			
POW	√			
CDM	√	√	√	
KNN	√	√		√

a chicken-and-egg flavor, because solving one relies on the other. In practice, we can only make some approximations to the weighting function in FRS.

3.2.1. The original space

The simplest method is to define the weighting function in the original feature space $\mathbb{R}^d$, and this method has the lowest computation cost. However, the weighting function in the original space and that in the FRS may be significantly different. The large-distance class pairs (small f_{ij}) in the original space may become small-distance pairs (large f_{ij}) in FRS. Therefore, the weighting function in the original space still suffers from the class separation problem, which will weaken the classification performance.

3.2.2. The low-dimensional space

To better approximate the weighting function in FRS, we can define weighting functions in the low-dimensional space. That means we first construct a weighting matrix in the original space $\mathbb{R}^d$ and learn a WFC transformation matrix $W \in \mathbb{R}^{d \times d'}$ to transform the data into $\mathbb{R}^{d'}$. After that, the weighting matrix is re-estimated in this low-dimensional space $\mathbb{R}^{d'}$ and is then incorporated into WFC to learn the transformation matrix again. The weighting matrix in the low-dimensional space reflects the class confusion relationship more accurately, and can achieve better performance than the weighting matrix in the original space. The cycle of learning $W \in \mathbb{R}^{d \times d'}$ and weighting matrix in $\mathbb{R}^{d'}$ can be repeated iteratively. However, the convergence is not guaranteed because the objective function optimized by this iterative procedure is not well-defined.

3.2.3. The fractional space

A useful iterative method to approximate the weighting function in FRS is the fractional-step dimensionality reduction [33]. The dimensionality is reduced from d to d' ($d' < d$) in small fractional steps, allowing for the relevant class pairs to be more correctly weighted. Denote by t the fractional step, the dimensionality reduction is performed in multiple steps iteratively as

$$\mathbb{R}^d \xrightarrow[\text{[WFC]}]{F} \mathbb{R}^{d-t} \xrightarrow[\text{[WFC]}]{F} \mathbb{R}^{d-2t} \cdots \xrightarrow[\text{[WFC]}]{F} \mathbb{R}^{d'}. \tag{25}$$

In each step, the weighting matrix F is estimated in the higher-dimensional input space, and then WFC is used to reduce the dimensionality by a small step t, after which the weighting matrix is re-estimated in the lower-dimensional output space. This procedure is repeated until the final dimensionality reaches d'. The fractional-step t can be very small. The step $t < 1$ means many steps are involved to reduce the dimensionality by 1 (from $\mathbb{R}^d$ to $\mathbb{R}^{d-1}$) (more details can be found in [33]). To alleviate the computation burden for large category applications, in this paper we only consider the fractional step t to be an integer (e.g. 1,5,10). By reducing the dimensionality with small fractional steps, the weighting matrix F can be estimated more reliably and closer to the weighting matrix in the FRS.

3.2.4. Comparison of weighting spaces

The three weighting spaces have increasing computational complexities, but lead to better approximations of the weighting function in the FRS. All the five weighting functions (FDA, aPAC, POW, CDM and KNN) can be defined in the three weighting spaces. If the weighting function is not changed for different weighting spaces (space invariant), we can simply define the weighting function in the original space which has the lowest computational complexity. We will show the improvements by considering weighting functions in different weighting spaces in Section 4.8. We also compare the space invariance property of different weighting functions in Section 4.9.

4. Evaluation of weighted Fisher criteria

We evaluated the weighted Fisher criteria (WFC) with five weighting functions (FDA, aPAC, POW, CDM and KNN) in three different weighting spaces (original space, low-dimensional space and fractional space) on a large scale 3,755-class Chinese handwriting dataset CASIA-HWDB1.1 [27].

4.1. Dataset

The CASIA-HWDB1.1 [27] is a new handwritten Chinese character database collected by the Institute of Automation of Chinese Academy of Sciences. This database contains samples from 300 writers (240 for training and 60 for testing). Each writer produced one sample for each of the 3,755 classes (GB2312-80 level-1 set), but a few mis-written samples were abandoned. Finally, 897,758 training samples and 223,991 test samples were obtained. For representing a character sample, features from gray-scale character images (background eliminated) were extracted using the normalization-cooperated gradient feature (NCGF) method [25]. The feature dimensionality is 512, representing the histogram of gradients extracted at 8 directions in 8×8 spatial grids. The feature data can be downloaded from our database webpage [1].

Chinese handwriting recognition is a challenging problem due to the large category (thousands of classes) and the many similar characters [14]. Some similar character pairs are shown in Fig. 8.

Table 3
Classification accuracies (%) of different dimensionality reduction models with the NCM classifier.

d'	60	70	80	90	100	110	120	130	140	150	160	170	180	Average
FDA	78.80	79.88	80.56	81.07	81.43	81.71	81.88	81.97	82.09	82.12	82.13	82.16	82.19	81.38
aPAC	78.84	79.92	80.58	81.08	81.41	81.74	81.89	81.96	82.05	82.12	82.13	82.13	82.16	81.39
aPAC-L	78.94	80.01	80.62	81.09	81.44	81.75	81.88	81.94	82.06	82.12	82.13	82.14	82.16	81.41
aPAC-F10	78.95	80.03	80.63	81.11	81.48	81.76	81.91	82.00	82.07	82.15	82.16	82.16	82.18	81.43
aPAC-F5	78.95	80.03	80.63	81.12	81.49	81.76	81.91	81.99	82.07	82.14	82.16	82.17	82.19	81.43
aPAC-F1	78.94	80.03	80.63	81.12	81.50	81.76	81.91	81.99	82.07	82.14	82.16	82.16	82.19	81.43
POW10	79.11	80.18	80.81	81.27	81.61	81.84	81.98	82.04	82.10	82.14	82.18	82.13	82.13	81.50
POW9	79.23	80.21	80.86	81.26	81.55	81.82	82.00	82.07	82.13	82.16	82.16	82.19	82.13	81.52
POW9-L	78.57	80.13	80.95	81.49	81.80	82.03	82.12	82.22	82.21	82.27	82.26	82.26	82.23	81.58
POW9-F10	79.75	80.58	81.12	81.52	81.79	81.97	82.11	82.18	82.19	82.24	82.23	82.24	82.18	81.70
POW9-F5	79.77	80.61	81.14	81.51	81.81	81.96	82.11	82.19	82.20	82.24	82.24	82.24	82.18	81.71
POW9-F1	79.77	80.62	81.15	81.52	81.81	81.98	82.12	82.18	82.19	82.24	82.23	82.24	82.18	81.71
POW8	79.13	80.19	80.85	81.19	81.58	81.80	81.95	82.03	82.13	82.15	82.15	82.15	82.15	81.50
POW7	79.02	80.13	80.82	81.20	81.54	81.77	81.92	82.02	82.11	82.13	82.15	82.15	82.15	81.47
CDM	79.20	80.21	80.77	81.25	81.52	81.73	81.89	82.04	82.05	82.09	82.05	82.12	82.16	81.47
CDM-L	79.66	80.52	81.01	81.35	81.62	81.84	81.97	82.09	82.12	82.12	82.12	82.13	82.14	81.59
KNN1	79.76	80.68	81.29	81.61	81.91	82.07	82.21	82.31	82.35	82.36	82.36	82.29	82.29	81.81
KNN5	80.30	81.20	81.67	82.02	82.17	82.32	82.46	82.43	82.44	82.49	**82.46**	82.37	**82.35**	82.05
KNN5-L	80.41	81.27	81.81	82.04	**82.29**	**82.38**	**82.50**	**82.48**	**82.49**	82.46	82.41	**82.40**	82.33	82.10
KNN5-F10	80.56	**81.37**	**81.82**	82.09	82.26	82.35	82.49	82.47	82.48	82.47	82.44	**82.40**	**82.35**	**82.12**
KNN5-F5	**80.57**	81.34	81.81	**82.11**	82.26	82.33	82.47	82.47	82.48	**82.50**	82.43	82.38	82.34	82.11
KNN5-F1	**80.57**	81.35	81.80	82.10	82.27	82.34	82.48	82.46	82.48	82.49	82.44	**82.40**	82.33	**82.12**
KNN10	80.31	81.12	81.62	81.95	82.15	82.29	82.42	82.41	82.46	82.45	82.38	82.34	**82.35**	82.02

In dimensionality reduction for Chinese handwriting recognition, the ordinary FDA suffers from the class separation problem that it will merge the similar character pairs that are close to each other in the original feature space. Therefore, the weighted Fisher criteria (WFC) are necessary and important to improve the performance for large category dimensionality reduction.

4.2. Classifiers

Two efficient and effective large-category classifiers are used for our evaluations.[2] The first classifier, denoted as nearest class mean (NCM), is based on the Euclidean distance

$$x \in \arg\min_{i=1}^{C}\{d_1(x,i) = \|x-\mu_i\|_2^2\}. \tag{26}$$

The second classifier is the quadratic discriminant function (QDF) derived from the Bayes decision theory under the assumption of Gaussian class-conditional distribution

$$x \in \arg\min_{i=1}^{C}\{d_2(x,i) = (x-\mu_i)^\top \Sigma_i^{-1}(x-\mu_i)+\log|\Sigma_i|\}, \tag{27}$$

where $\mu_i \in \mathbb{R}^d$ is the mean vector and $\Sigma_i \in \mathbb{R}^{d\times d}$ is the covariance matrix of class i. To efficiently compute Σ_i^{-1} in $d_2(x,i)$, we use the modified quadratic discriminant function (MQDF) [20] which is the state-of-the-art classifier in Chinese handwriting recognition. The MQDF replaces the small eigenvalues of Σ_i with a constant, so that only the principal eigenvectors are needed in computing the quadratic distance, and improved classification performance is obtained. The MQDF uses $d_3(x,i)$ to replace $d_2(x,i)$, where

$$d_3(x,i) = \sum_{j=1}^{k}\frac{1}{\lambda_{ij}}[(x-\mu_i)^\top\phi_{ij}]^2 + \frac{1}{\delta_i}\left\{\|x-\mu_i\|^2 - \sum_{j=1}^{k}[(x-\mu_i)^\top\phi_{ij}]^2\right\} + \sum_{j=1}^{k}\log\lambda_{ij} + (d-k)\log\delta_i, \tag{28}$$

where $\lambda_{ij} \in \mathbb{R}^+$ and $\phi_{ij} \in \mathbb{R}^d$, $j=1,\ldots,d$, are respectively the eigenvalues (sorted in non-ascending order) and their corresponding eigenvectors of Σ_i, and k is the number of principal eigenvectors. The minor eigenvalues $\lambda_{i,k+1}\cdots\lambda_{i,d}$ are replaced with a constant δ_i. We set δ_i to be common for all the classes and selected its value by cross validation on the training dataset. The number of principal components used in MQDF was empirically set to be $k=50$ for all the methods in our experiments.

In the following sections, we use both NCM (26) and MQDF (28) to evaluate the classification performance in the dimensionality reduced spaces of WFC. MQDF is a state-of-the-art classifier in Chinese handwriting recognition over the past 25 years. MQDF can give much higher classification accuracy than NCM. However, for MQDF the memory requirement is much heavier and the testing speed is much slower than NCM. For example, in 160-dimensional space, the memory requirement is 117MB for MQDF and 2.29MB for NCM, and the classification speed is 12.25 millisecond/character for MQDF (speeded up via candidate selection by NCM) and 1.85 millisecond/character for NCM. Therefore, using NCM and MQDF for comparing different dimensionality reduction models is a widely used strategy for Chinese handwriting recognition [15,42,30].

4.3. Experimental settings

The five weighting functions in three different weighting spaces are compared according to the classification accuracy in the reduced low-dimensional space from $d=512$ to $d'=60,70,\ldots,180$.

For the POW method, we test powers $m=3,4,\ldots,12$ for Eq. (22) and report the best performance on the test dataset. For the CDM method in Eq. (23), we partition the training set randomly into two subsets: using 3/4 for training the basic classifier and 1/4 for estimating the confusion matrix. For the KNN method in Eq. (24), we evaluate the performance by varying $k=1,5,10$. For the fractional-step dimensionality reduction (25), we compare the performance with fractional step $t=1,5,10$. The algorithms were programmed in C++ and executed on a PC (CPU: Intel Dual E8400 3.0 GHz, RAM: 2 GB).

4.4. Experimental results

The experimental results are shown in Table 3 for the NCM classifier and Table 4 for the MQDF classifier. The "POW10, POW9,

[2] Other classifiers such as the nearest neighbor (NN) classifier and support vector machines (SVM) are too expensive for large-category problems.

Table 4
Classification accuracies (%) of different dimensionality reduction models with the MQDF classifier.

d'	60	70	80	90	100	110	120	130	140	150	160	170	180	Average
FDA	86.35	87.42	88.14	88.59	88.87	89.10	89.26	89.40	89.47	89.52	89.53	89.51	89.51	88.82
aPAC	86.33	87.42	88.13	88.61	88.87	89.10	89.26	89.36	89.46	89.50	89.49	89.48	89.50	88.81
aPAC-L	86.39	87.50	88.15	88.63	88.89	89.10	89.27	89.40	89.47	89.49	89.50	89.50	89.50	88.83
aPAC-F10	86.43	87.54	88.22	88.68	88.94	89.17	89.32	89.42	89.53	89.54	89.56	89.52	89.51	88.88
aPAC-F5	86.43	87.54	88.22	88.68	88.94	89.17	89.32	89.42	89.52	89.54	89.56	89.52	89.51	88.87
aPAC-F1	86.43	87.54	88.21	88.69	88.94	89.17	89.31	89.42	89.53	89.54	89.56	89.52	89.50	88.87
POW10	86.46	87.49	88.20	88.68	88.95	89.11	89.26	89.35	89.39	89.45	89.48	89.47	89.47	88.83
POW9	86.55	87.61	88.29	88.75	89.00	89.20	89.34	89.44	89.47	89.48	89.54	89.52	89.53	88.90
POW9-L	85.71	87.29	88.15	88.69	89.02	89.28	89.44	89.49	89.57	89.57	89.57	89.54	89.53	88.83
POW9-F10	87.04	87.92	88.52	88.93	89.18	89.37	89.51	89.57	89.61	89.58	89.59	89.60	89.59	89.08
POW9-F5	87.04	87.93	88.52	88.92	89.19	89.35	89.51	89.58	89.60	89.58	89.59	89.62	89.59	89.08
POW9-F1	87.05	87.95	88.56	88.93	89.20	89.36	89.51	89.58	89.62	89.58	89.59	89.62	89.58	89.09
POW8	86.53	87.59	88.26	88.72	88.98	89.19	89.31	89.43	89.48	89.53	89.52	89.54	89.53	88.89
POW7	86.46	87.53	88.24	88.70	88.93	89.17	89.32	89.40	89.50	89.52	89.53	89.53	89.49	88.87
CDM	86.81	87.83	88.45	88.86	89.14	89.33	89.40	89.48	89.51	89.54	89.52	89.53	89.49	88.99
CDM-L	87.10	87.96	88.55	88.93	89.15	89.34	89.40	89.47	89.53	89.54	89.52	89.49	89.47	89.03
KNN1	87.03	87.93	88.53	88.94	89.19	89.30	89.40	89.49	89.57	89.59	89.61	89.55	89.50	89.05
KNN5	87.50	88.35	88.85	**89.24**	89.43	**89.59**	89.66	89.73	89.73	**89.76**	89.73	89.71	**89.71**	89.31
KNN5-L	87.46	88.35	88.88	89.23	89.49	89.55	**89.70**	**89.74**	89.75	89.73	89.73	89.74	89.66	89.31
KNN5-F10	87.63	**88.37**	88.89	89.23	**89.50**	**89.59**	89.66	**89.74**	89.77	89.74	89.76	**89.78**	**89.71**	**89.34**
KNN5-F5	**87.64**	**88.37**	88.89	89.22	89.49	89.58	89.68	89.73	**89.78**	89.74	**89.77**	89.74	**89.71**	89.33
KNN5-F1	87.63	88.36	**88.90**	89.19	89.48	89.57	89.69	89.72	89.76	89.73	89.76	89.74	89.70	89.33
KNN10	87.55	88.32	88.84	89.17	89.46	89.55	89.65	89.71	89.70	89.71	89.70	89.68	89.68	89.29

POW8, POW7" means the POW weighting function (22) with $m=10,9,8,7$, respectively. The "KNN1, KNN5, KNN10" means KNN weighting function (24) with $k=1,5,10$, respectively. For different weighting spaces, we use "xxx" to denote the weighting function in the original space, "xxx-L" to denote the low-dimensional space, and "xxx-F10, xxx-F5, xxx-F1" to denote the fractional spaces with fractional step $t=10,5,1$ respectively. For example, "KNN5-F10" means the KNN weighting function with $k=5$ in the fractional space with $t=10$. We did not consider the CDM weighting function in the fractional spaces, because in each step CDM needs to train a classifier and estimate a confusion matrix, which is very time-consuming when combined with fractional steps.

Our experiments were conducted on the standard training and testing sets of CASIA-HWDB1.1 database, and therefore our results are not comparable with the results of the Chinese handwriting recognition competition in ICDAR2011 [29] where the training data were open. To further improve the performance of Chinese handwriting recognition, we can (i) enlarge the training dataset [28], (ii) use discriminatively trained classifiers [26], (iii) adopt some perturbation or distortion based method to produce multiple new patterns for a testing pattern and then combine multiple decisions to boost the accuracy [37], (iv) apply the convolutional neural network to automatically learn the features from the data [8], and also (v) adapt the classifier to the unique handwriting style of a particular writer [49]. Because our purpose is to evaluate different dimensionality reduction models, we did not use these strategies to further improve the classification accuracy.

From the results in Tables 3 and 4, we can see that the class separation problem does occur in Chinese handwriting recognition. Compared with the baseline FDA model, the weighted Fisher criteria (WFC) can improve the classification accuracies consistently for all the reduced subspaces and for both the NCM and MQDF classifiers. The improvement is significant especially when the reduced dimensionality is low. For example, at a dimension of 60 for MQDF classifier, the accuracy is improved from 86.35% for FDA to 87.64% for KNN5-F5. On the other hand, at a dimension of 180, the accuracy is only improved from 89.51% for FDA to 89.71% for KNN5-F5. The reason is that, with the increase of dimensionality, the classification performance will become saturated, and the differences of the performance between different dimensionality reduction models will become smaller (this can be verified from the results in Tables 3 and 4). However, improvement of the classification accuracy at low dimensional spaces is of practical value, e.g., for embedding the classifiers into some handheld devices (e.g. mobile phone and tablet computer) which requires the memory requirement to be as low as possible and also the computation speed to be as fast as possible (in such applications, the dimensionality can be as low as 30). Besides the accuracies shown in Tables 3 and 4, in the following we also make a fair and meaningful comparison of their average ranks to show the statistical significance of different models.

4.5. Statistical significance

Friedman test is a widely used statistical test for comparing more than two algorithms over multiple evaluations [10]. Suppose we have k algorithms evaluated for N times. Let r_i^j be the rank of the j-th algorithm in the i-th evaluation.[3] The Friedman test compares the average ranks of algorithms $R_j=\frac{1}{N}\sum_{i=1}^{N} r_i^j$. The null-hypothesis states that all the algorithms are equivalent, and so, their ranks $R_j, j=1,\ldots,k$ should be equal. If the null hypothesis is rejected at 0.05 significance level[4], we can proceed with a post-hoc test (the *Nemenyi test*) to find out which algorithms significantly differ. Specifically, the performance of two algorithms are significantly different if their average ranks differ by at least the *critical difference* (CD) [10].

In Tables 3 and 4, we have a total of 23 models with 26 evaluations. Fig. 3 shows the CD diagram for the 23 models, where the average rank of each compared model is marked along the axis. The axis is turned so that the lowest (best) ranks are to the right. Groups of models that are not significantly different are connected with a thick line. The critical difference ($CD=6.802$ at 0.05 significance level) is also shown above the axis. If the difference of the average ranks of two models $R_A-R_B\geq 6.802$

[3] In case of ties, average ranks are assigned, e.g. the ranks of {80%,90%,70%,70%,60%} are {2,1,3.5,3.5,5}.

[4] If we reject the null hypothesis, we may make a mistake with probability 0.05.

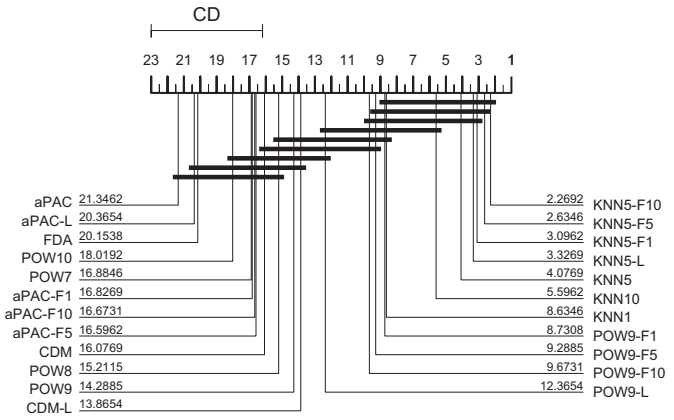


Fig. 3. The critical difference (CD) diagram of different methods.

(CD), we can conclude that model "B" significantly outperforms model "A".

4.6. Comparison of five weighting functions

From the results in Tables 3 and 4, we can find that the best result for the POW method is achieved when $m=9$; for the KNN method, $k=5$ gives the best performance. From Fig. 3, we calculate the differences of average ranks of the five weighting functions as:

$$\text{aPAC} \overset{1.1924}{\longrightarrow} \text{FDA} \overset{4.0769}{\longrightarrow} \text{CDM} \overset{1.7884}{\longrightarrow} \text{POW9} \overset{10.2116}{\longrightarrow} \text{KNN5}. \tag{29}$$

The models on the right side are better than the ones on the left. We observe that the differences between aPAC and FDA, also between CDM and POW9 are not significant. The CDM, POW9 and KNN5 have significantly higher classification accuracies than FDA, especially when the reduced dimensionality is low. For example, when $d'=60$ in Table 3, the accuracies are 78.80%, 79.20%, 79.23% and 80.30% for FDA, CDM, POW9 and KNN5, respectively. Furthermore, the KNN5 weighting function significantly outperforms all the other models, because the differences of the average ranks between KNN5 and the other models are much larger than 6.802. Compared with the baseline FDA model, KNN5 significantly improves the classification accuracies as shown in Fig. 4.

4.7. Comparison of computational complexity

The computational cost of WFC based dimensionality reduction includes three parts[5]: (i) the computation of the weighting matrix, (ii) the construction of the scatter matrix $\widehat{S}_b$ in Eq. (15), and (iii) the eigen-decomposition of $\widehat{S}_b$. The comparison of the computation times of the five weighting functions are shown in Fig. 5. The training time of CDM is based on the NCM classifier. For the MQDF classifier, CDM will have even much longer training time.

We find that FDA has the lowest running time. This is because FDA uses a constant weighting function $f_{ij}=1,\forall i,j$ (no need to compute the weighting matrix). Furthermore, the scatter matrix of FDA can be computed efficiently as:

$$\widehat{S}_b = \sum_{i,j=1}^{C} p_i p_j (\widehat{\mu}_i - \widehat{\mu}_j)(\widehat{\mu}_i - \widehat{\mu}_j)^\top = 2\sum_{i=1}^{C} p_i (\widehat{\mu}_i - \widehat{\mu_0})(\widehat{\mu}_i - \widehat{\mu_0})^\top, \tag{30}$$

[5] Because the whitening transformation is a common pre-processing step for all the weighting functions, we omit the computation time of whitening here.

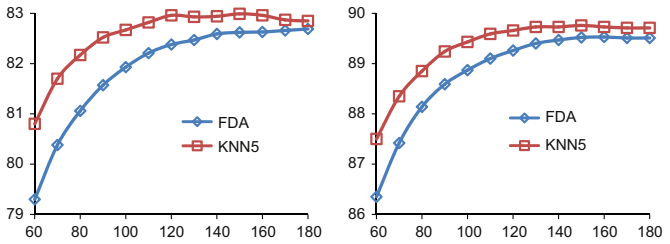
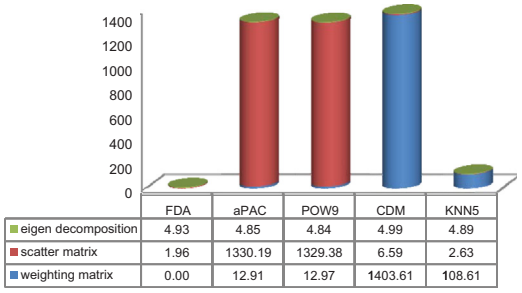


Fig. 4. Comparison of the classification accuracies (%) in different reduced spaces for KNN5 and FDA with the NCM (left) and MQDF (right) classifiers.



	FDA	aPAC	POW9	CDM	KNN5
eigen decomposition	4.93	4.85	4.84	4.99	4.89
scatter matrix	1.96	1330.19	1329.38	6.59	2.63
weighting matrix	0.00	12.91	12.97	1403.61	108.61

Fig. 5. Training times (seconds) for different weighting functions ($d'=160$).

where $\widehat{\mu_0}=\sum_{j=1}^{C} p_j\widehat{\mu}_j$. While for the other weighting functions, the scatter matrix can only be computed in a pairwise manner:

$$\widehat{S}_b = \sum_{i,j=1}^{C} f_{ij} p_i p_j (\widehat{\mu}_i - \widehat{\mu}_j)(\widehat{\mu}_i - \widehat{\mu}_j)^\top = \sum_{i=1}^{C}\sum_{j=i+1}^{C} (f_{ij}+f_{ji}) p_i p_j (\widehat{\mu}_i - \widehat{\mu}_j)(\widehat{\mu}_i - \widehat{\mu}_j)^\top. \tag{31}$$

This is extremely time-consuming when C is large. For example, for the aPAC and POW9 methods in Fig. 5, the computation time of the scatter matrix is nearly 75 times that of the other steps combined. Therefore, for the weighted Fisher criteria (except FDA), the locality (sparsity) of the weighting matrix is very important for the reduction of the computation cost (as it eliminates the need to calculate the terms in $\widehat{S}_b$ when $f_{ij}+f_{ji}=0$). As discussed in Section 3.1.6, the weighting matrices of CDM and KNN are very sparse (Table 2). Therefore, the computation times of the scatter matrices for CDM and KNN are significantly reduced as shown in Fig. 5. However, for the CDM method, a very time-consuming process is the estimation of the confusion matrix which needs to train a classifier and evaluate it on a validation dataset. Considering the overall computation time, with the exception of the baseline FDA model, the KNN method has the lowest computational complexity due to the fast construction and sparsity of the weighting matrix.

4.8. Comparison of three weighting spaces

From Fig. 3, we can find that: aPAC-L outperforms aPAC; CDM-L outperforms CDM; POW9-L outperforms POW9; and KNN5-L outperforms KNN5. This indicates that the weighting function in the low-dimensional space can achieve better performance than that in the original feature space. Take the CDM as an example, when $d'=60$, CDM-L improves the classification accuracy from 79.20% to 79.66% in Table 3 and from 86.81% to 87.10% in Table 4. This is because the confusion matrix estimated in the low-dimensional space is more relevant to the real confusion information in the final reduced space.

The fractional-step spaces can further improve the classification performance compared with the low-dimensional space.

The average ranks (Fig. 3) are improved by:

$$\text{aPAC}\xrightarrow{0.9808}\text{aPAC-L}\xrightarrow{3.7692}\text{aPAC-F5},$$

$$\text{POW9}\xrightarrow{1.9231}\text{POW9-L}\xrightarrow{3.6346}\text{POW9-F1},$$

$$\text{KNN5}\xrightarrow{0.7500}\text{KNN5-L}\xrightarrow{1.0577}\text{KNN5-F10}. \quad (32)$$

Take the POW9 model as an example, the comparison of the original space and fractional-step space is shown in Fig. 6. The improvements are not significant compared with the improvements caused by different weighting functions (29), however.

For different fractional steps, we can find that the performance is ordered as:

$$\text{aPAC-F1} < \text{aPAC-F10} < \text{aPAC-F5},$$

$$\text{POW9-F10} < \text{POW9-F5} < \text{POW9-F1},$$

$$\text{KNN5-F1} < \text{KNN5-F5} < \text{KNN5-F10}. \quad (33)$$

This indicates that small fractional steps do not always result in better classification performance.

We also compare the training times for different weighting spaces in Fig. 7. We find that when the fractional step t decreases, the computational complexity increases dramatically. Therefore, in practice we should choose large fractional steps, considering the minor improvements produced by fractional-step spaces and the dramatically increased computation.

4.9. Space invariance analysis

In this section, we check the property of space invariance for different weighting functions. Given a weighting matrix $F=\{f_{ij}\}\in\mathbb{R}^{C\times C}$, we define the normalization of F as $\widehat{F}$:

$$\widehat{F}=\left\{\widehat{f_{ij}}=\frac{f_{ij}}{\sum_{i,j=1}^{C}f_{ij}}\right\}\in\mathbb{R}^{C\times C}. \quad (34)$$

This is to avoid the influence of scale change and will not affect the WFC criterion (13). Then we define the difference between two weighting matrices A and B as:

$$\text{diff1}=\sum_{i,j=1}^{C}|\widehat{A_{ij}}-\widehat{B_{ij}}|,$$

$$\text{diff2}=\frac{1}{C^2}\sum_{i,j=1}^{C}\mathbb{I}(\widehat{A_{ij}}\neq\widehat{B_{ij}}), \quad (35)$$

where $\mathbb{I}(a\neq b)=1$ if $a\neq b$ and 0 otherwise. The "diff1" measures the absolute error and "diff2" counts the number of different elements in A and B.

We calculate the difference for the weighting functions between the original space and final reduced space. The comparison is shown in Table 5. The confusion matrix used by CDM is classifier dependent, hence we show two results of CDM for the classifiers NCM and MQDF respectively. The FDA weighting matrix is absolutely space invariant. The aPAC model has low absolute error ("diff1") but the elements are mostly changed in the original and final reduced space ("diff2"). Considering both "diff1" and "diff2", the KNN model is nearly space invariant, i.e., the KNN relationship in different spaces is nearly unchanged. This explains why the fractional-step space achieves the lowest improvements on KNN (32).

4.10. Similar character analysis

In this section, we compare the classification performance on similar characters in the dimensionality reduced space. We define the confusion probability (CP) of two classes A and B as:

$$CP_{A,B}=\frac{1}{2}\left(\frac{N_{A\to B}}{N_A}+\frac{N_{B\to A}}{N_B}\right), \quad (36)$$

where N_A is the number of samples in class A, and $N_{A\to B}$ is the number of samples of class A that are classified into class B. The confusion probabilities of some similar characters in the reduced spaces of FDA and KNN5 ($d'=60$ with NCM classifier) are shown in Fig. 8. We can find that: (1) for most similar characters, the confusion probabilities of KNN5 are significantly lower than that of FDA; (2) however, for some particular similar characters, the confusion probabilities of KNN5 are higher than that of FDA. On the testing data of CASIA-HWDB1.1, there are totally 8279 class pairs with confusion probabilities higher than 0.01 (for either FDA or

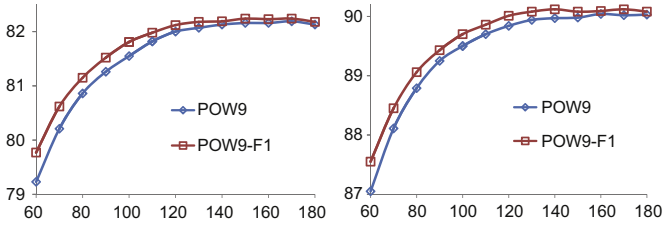


Fig. 6. Comparison of classification accuracies (%) in different reduced spaces for POW9 and POW9-F1 with the NCM (left) and MQDF (right) classifiers.

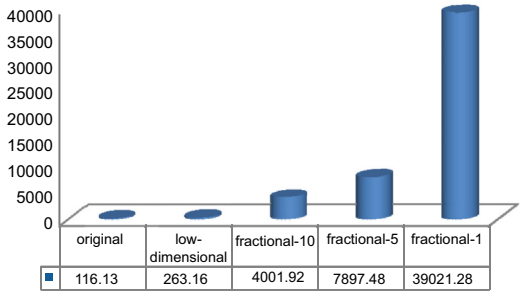


Fig. 7. Training times (seconds) for the KNN5 model in different weighting spaces ($d'=160$).

Table 5
Differences of weighting functions between original ($d=512$) and final reduced spaces ($d'=160$).

	diff1	diff2
FDA	0.0000	0.0000
aPAC	0.0286	0.9997
POW9	0.5201	0.3845
CDM (NCM)	0.7283	0.0028
CDM (MQDF)	1.0106	0.0022
KNN5	0.2585	0.0003

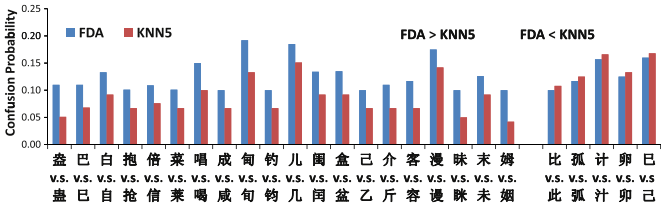


Fig. 8. Confusion probabilities of some similar characters for FDA and KNN5.

KNN5). Among the 8279 class pairs, KNN5 outperforms FDA on 3755 class pairs, FDA outperforms KNN5 on 2650 class pairs, and for the remaining class pairs FDA and KNN5 have the same performance. These results confirm the advantages of KNN5 in improving the overall classification performance for similar characters.

5. Extension of weighted Fisher criteria from class level to sample level

The weighted Fisher criteria (WFC) only use the mean vector and covariance matrix of each class, i.e., consider the weighted scatter matrix at *class level* as Eq. (15). Therefore, WFC can be viewed as a parametric feature extraction method [12]. Nonparametric discriminant analysis (NDA) methods [13,22] extend the scatter matrix to the *sample level*, wherein different samples can be weighted appropriately in the scatter matrix according to their nearness to the decision boundaries in the flavor of decision boundary based feature extraction [23]. Considering that weighting at the sample level exploits more discriminative information than weighting at the class level, we also extend WFC to the sample level.

5.1. SKNN: Sample level KNN method

Because the KNN weighting function performs best in the previous evaluations, we extend the KNN method to the sample level. Given a training dataset $\{x_i,y_i\}_{i=1}^N$ where $x_i \in \mathbb{R}^d$ and $y_i \in \{1,\ldots,C\}$. The whitening transformation W_{whiten} is defined in (7), and the whitened class mean $\widehat{\mu}_i$ is defined in (16). The between-class scatter matrix is now extended to the sample level as:

$$\tilde{S}_b = \sum_{i=1}^{N}\sum_{j=1}^{C} f_{ij}(\widehat{x}_i-\widehat{\mu}_j)(\widehat{x}_i-\widehat{\mu}_j)^\top, \tag{37}$$

where $\widehat{x}_i = W_{\text{whiten}}^\top x_i$. The sample level weighting function $F=\{f_{ij}\} \in \mathbb{R}^{N\times C}$ is now defined as:

$$f_{ij} = \begin{cases} 1, & \text{if } \widehat{\mu}_j \in \text{KNN}(\widehat{x}_i,y_i) \\ 0, & \text{otherwise} \end{cases} \tag{38}$$

where $\text{KNN}(\widehat{x}_i,y_i)$ denotes the k nearest neighbors of $\widehat{x}_i$ in $\{\widehat{\mu_1},\widehat{\mu_2},\ldots,\widehat{\mu_{y_i-1}},\widehat{\mu_{y_i+1}},\ldots,\widehat{\mu_C}\}$. The scatter matrix $\tilde{S}_b$ is based on the sample level KNN method, therefore we call this method SKNN. Denote the columns of $W_{\text{SKNN}} \in \mathbb{R}^{d\times d'}$ to be the d' eigenvectors corresponding to the d' largest eigenvalues of $\tilde{S}_b$, the final dimensionality reduction matrix is: $W_{\text{final}} = W_{\text{whiten}}W_{\text{SKNN}} \in \mathbb{R}^{d\times d'}$.

SKNN is a nonparametric extension of the KNN method to the sample level. Different samples can have different nearest neighbors of class means, and therefore SKNN can capture much more information about the decision boundary. SKNN can overcome the class separation problem because each sample is only connected to its k nearest neighbors. SKNN can also alleviate the heteroscedastic problem, since $\tilde{S}_b$ is computed from all the samples other than the class means. By computing the between-class scatter matrix only from the class means as in parametric WFC (15), the covariance matrices of each class are only used in whitening (i.e. S_w) which is based on the homoscedastic assumption. Contrarily, by computing the between-class scatter matrix from all the samples in SKNN, the information of the covariance matrices can be implicitly captured in the second step of eigen-decomposition of $\tilde{S}_b$. Furthermore, SKNN can partially solve the multi-modal distribution problem, because each class is described not only by its class mean but also the entire training sample set of this class. In summary, SKNN can capture much more information of the decision boundary, solve the class separation problem, and also alleviate the heteroscedastic and multi-modal problems. Therefore, SKNN is expected to achieve much better performance than the other models.

In the next sections, we compare the performance of SKNN with other methods which have shown high performance for Chinese handwriting recognition in the recent literature.

5.2. Other dimensionality reduction methods for Chinese handwriting recognition

The locally linear discriminant analysis (LLDA) recently proposed by Gao et al. [15] uses three strategies to improve the classification performance: (1) partition each class into several clusters; (2) find the nearest neighboring clusters from the remaining classes for each cluster of one class, and use the corresponding cluster means to compute the between-class scatter matrix; and (3) apply feature vector normalization to further improve the performance. The LLDA can solve the class separation problem and also the multi-modal sample distribution problem, and hence has shown better performance than the traditional FDA in Chinese handwriting recognition [15].

The neighbor class linear discriminant analysis (NCLDA) was recently proposed by Wang et al. [42] to solve the class separation problem. NCLDA re-defines the between-class scatter matrix as $S_b = \sum_{i=1}^{C} p_i(\widehat{\mu}_i - \frac{1}{k}\sum_{j=1}^{k}\widehat{\mu_{ij}})(\widehat{\mu}_i - \frac{1}{k}\sum_{j=1}^{k}\widehat{\mu_{ij}})^\top$, where $\widehat{\mu_{ij}}$ is the j-th nearest neighbor of $\widehat{\mu}_i$ from the remaining classes. When $k=C$, NCLDA is equivalent to the traditional FDA. By setting a small value of k, NCLDA can solve the class separation problem. NCLDA is very similar to the KNN based weighted Fisher criterion. The difference lies in that: the KNN based weighted Fisher criterion maximizes the distance between each class and each of its nearest neighbors; while the NCLDA maximizes the distance between each class and the mean of its nearest neighbors. Utilizing the mean of the nearest neighbors may lose some important discriminative information. For example, consider three classes located at A: $(-1,0)$, B: $(0,0)$, and C: $(1,0)$. Maximizing the distances of $d(B,C)+d(B,A)$ will find the x-axis as the projection direction. Contrarily, maximizing the distance of $d(B,0.5(A+C))$ cannot produce any meaningful results. Therefore, the KNN based weighted Fisher criterion should be more robust than NCLDA theoretically.

Another limitation of FDA is the heteroscedastic problem, i.e., the covariance matrices are not the same for different classes, which breaks the homoscedastic assumption of FDA. Loog and Duin [31] proposed a heteroscedastic extension of FDA based on the Chernoff criterion. However, for large category problems, the pairwise-class calculation scheme used in the Chernoff criterion is computationally formidable. To solve this problem, Liu and Ding [30] proposed a heteroscedastic linear discriminant analysis (HLDA) scheme using the Mahalanobis criterion to replace the Chernoff criterion. This HLDA was shown to be more efficient than the Chernoff criterion and computationally feasible for large category problems, and has achieved promising performance in Chinese handwriting recognition.

5.3. Performance evaluation

In this section, we compare the performance of LLDA, NCLDA, HLDA with KNN based weighted Fisher criterion and the sample level extension of KNN (SKNN) on the CASIA-HWDB1.1 dataset [1] (see Section 4.1). For LLDA, we used the same parameter settings (the numbers of clusters and neighbors) reported by [15].[6] For

[6] In [15], the evaluations were not conducted on the standard training and testing partition of CASIA-HWDB1.1 [1], and the accuracies in [15] are much lower than our results.

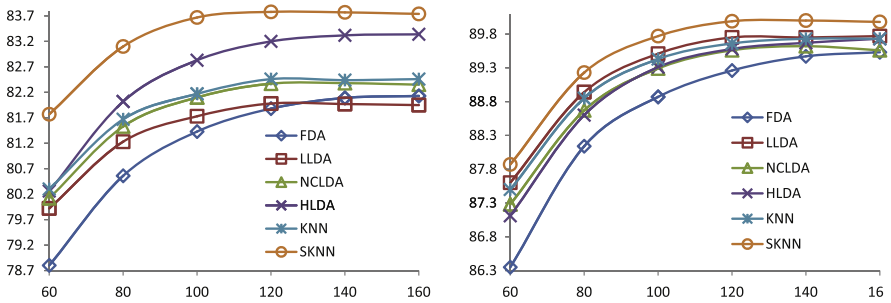


Fig. 9. Comparison of the classification accuracies (%) in different reduced spaces for FDA, LLDA, NCLDA, HLDA, KNN, and SKNN with the NCM (left) and MQDF (right) classifiers.

NCLDA, KNN, and SKNN, the number of nearest neighbors was set to be 5 for fair comparison.

The experimental results are shown in Fig. 9. From these results we can find that: (1) All the methods (LLDA, NCLDA, HLDA, KNN, and SKNN) can improve the classification accuracy compared with the traditional FDA. This is because FDA has the limitations of class separation problem and homoscedastic assumption. (2) KNN outperforms NCLDA for both NCM and MQDF classifiers. This indicates that maximizing the sum of the distances between each class and its nearest neighbor classes is more promising than maximizing the distance between each class and the mean of its nearest neighbor classes. (3) LLDA achieves good performance for MQDF but low performance for NCM. This identifies that partitioning each class into multiple clusters is useful for MQDF to capture the distributions of different classes. However, this strategy does not work for the nearest class mean (NCM) classifier which only uses one prototype (class mean) for each class. (4) HLDA achieves good performance for NCM but low performance for MQDF. This is because the NCM assumes Gaussian distribution with identity covariance matrices for each class. HLDA exploits the heteroscedastic information, and thus can improve the accuracy significantly for NCM. However, for the MQDF classifier, the differences of covariance matrices has already been taken into consideration, and in this case the main remaining issue is the class separation problem. Therefore, HLDA does not bring much gain to the MQDF classifier. (5) SKNN achieves the best performance for both the NCM and MQDF classifiers on all the reduced subspaces consistently. SKNN also outperforms the KNN method significantly. This demonstrates that the between-class scatter matrix calculated at sample level can capture much more discriminative information and is therefore much more accurate and robust than other approaches.

We also compare the training complexities of different models in Fig. 10. We can find that: (1) FDA, KNN, and NCLDA have lower training complexities, because they only use the class-wise means to calculate the between-class scatter matrix. (2) LLDA has moderate training complexity as it partitions each class into multiple clusters. (3) HLDA has the highest training complexity, because in the computation of the between-class scatter matrix, the inverse matrix operation is required for each class. (4) SKNN also has a high training complexity, because the complexity of sample level nearest neighbor searching is linearly dependent on the number of training samples, the number of classes and the number of original dimensionality. Considering the significant accuracy improvement brought by SKNN, the increased training complexity is worthwhile.

All the results reported here can be exactly repeated with the feature data released at [1], and we hope this can be used as a benchmark for comparing different dimensionality reduction methods on large category problems.

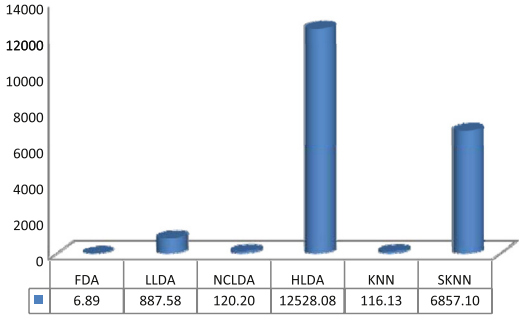


Fig. 10. Training times (seconds) for FDA, LLDA, NCLDA, HLDA, KNN, and SKNN ($d' = 160$).

6. Conclusion

In this paper, we investigate the weighted Fisher criteria (WFC) for solving the class separation problem in large category dimensionality reduction. The objective of WFC is to maximize the sum of weighted pairwise distances. By setting larger weights for the most confusable class pairs, WFC can improve the class separability in the reduced space. We evaluated five weighting functions (FDA, aPAC, POW, CDM and KNN) in three different weighting spaces (original space, low-dimensional space and fractional space) on a large scale 3,755-class Chinese handwriting dataset. The KNN weighting function achieves significantly better classification performance than the other weighting functions. Due to the sparsity and fast construction of the weighting matrix, the KNN method also has the lowest training complexity against the other weighting functions (except FDA). Different weighting spaces can improve the performance slightly with the cost of dramatically longer running time. It is also revealed that the KNN weighting matrix (KNN relationship between different classes) is nearly space invariant. Therefore, in practice, the KNN weighting function in the original space is the most efficient and effective model for large category dimensionality reduction.

We also extend the KNN based weighted Fisher criterion from class level to sample level. The sample level KNN (SKNN) is a nonparametric method which can capture much more information about the decision boundary, solve the class separation problem, and also alleviate the heteroscedastic and multi-modal problems. Experimental results identify that SKNN can outperform the locally linear discriminant analysis (LLDA) (proposed for solving class separation and multi-modal problems), neighbor class linear discriminant analysis (NCLDA) (proposed for solving class separation problem), and heteroscedastic linear discriminant analysis (HLDA) (proposed for solving heteroscedastic problem) for Chinese handwriting recognition.

Conflict of interest statement

None of declared.

Acknowledgments

We thank Prof. Louisa Lam for the proofreading of the manuscript. This work was supported in part by the National Basic Research Program of China (973 Program) Grant 2012CB316302, the National Natural Science Foundation of China (NSFC) Grants 60825301 and 60933010, and the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant XDA06030300).

References

[1] 〈http://www.nlpr.ia.ac.cn/databases/handwriting/Download.html〉.
[2] K.T. Abou-Moustafa, F. De La Torre, F.P. Ferrie, Pareto discriminant analysis, Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2010, pp. 3602–3609.
[3] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, Proc. Advances in Neural Information Processing Systems 14 (2001) 585–591.
[4] W. Bian, D. Tao, Harmonic mean for subspace selection, Proc. Int'l Conf. Pattern Recognition, 2008.
[5] W. Bian, D. Tao, Max-Min distance analysis by using sequential SDP relaxation for dimension reduction, IEEE Trans. Pattern Analysis and Machine Intelligence 33 (5) (2011) 1037–1050.
[6] E. Bingham, H. Mannila, Random projection in dimensionality reduction: applications to image and text data, Proc. Int'l Conf. Knowledge Discovery and Data Mining, 2001, pp. 245–250.
[7] H. Cevikalp, M. Neamtu, M. Wilkes, A. Barkana, Discriminative common vectors for face recognition, IEEE Trans. Pattern Analysis and Machine Intelligence 27 (1) (2005) 4–13.
[8] D. Ciresan, U. Meier, J. Schmidhuber, Multi-column deep neural networks for image classification, Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2012, pp. 3642–3649.
[9] S. Dasgupta, Experiments with random projection, Proc. Uncertainty in Artificial Intelligence, 2000, pp. 143–151.
[10] J. Demšar, Statistical comparisons of classifiers over multiple data sets, Journal of Machine Learning Research 7 (2006) 1–30.
[11] R.A. Fisher, The use of multiple measurements in taxonomic problems, Annals of Eugenics 7 (2) (1936) 179–188.
[12] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, 1990.
[13] K. Fukunaga, J.M. Mantock, Nonparametric discriminant analysis, IEEE Trans. Pattern Analysis and Machine Intelligence (6) (1983) 671–678.
[14] T.-F. Gao, C.-L. Liu, High accuracy handwritten Chinese character recognition using LDA-based compound distances, Pattern Recognition 41 (11) (2008) 3442–3451.
[15] X. Gao, J. Guo, L. Jin, Dimensionality reduction by locally linear discriminant analysis for handwritten Chinese character recognition, IEICE Trans. Information and Systems 95 (10) (2012) 2533–2543.
[16] X. He, P. Niyogi, Locality preserving projections, Proc. Advances in Neural Information Processing Systems, 2004.
[17] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.
[18] A. Hyvärinen, E. Oja, Independent component analysis: algorithms and applications, Neural Networks 13 (4-5) (2000) 411–430.
[19] I.T. Jolliffe, Principal Component Analysis, Springer-verlag, 1986.
[20] F. Kimura, K. Takashina, S. Tsuruoka, Y. Miyake, Modified quadratic discriminant functions and the application to Chinese character recognition, IEEE Trans. Pattern Analysis and Machine Intelligence 9 (1) (1987) 149–153.
[21] F. Kimura, T. Wakabayashi, Y. Miyake, On feature extraction for limited class problem, Proc. Int'l Conf. Pattern Recognition, 1996, pp. 191–194.
[22] B.-C. Kuo, D.A. Landgrebe, Nonparametric weighted feature extraction for classification, IEEE Trans. Geoscience and Remote Sensing 42 (5) (2004) 1096–1105.
[23] C. Lee, D.A. Landgrebe, Feature extraction based on decision boundaries, IEEE Trans. Pattern Analysis and Machine Intelligence 15 (4) (1993) 388–400.
[24] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (6755) (1999) 788–791.
[25] C.-L. Liu, Normalization-cooperated gradient feature extraction for handwritten character recognition, IEEE Trans. Pattern Analysis and Machine Intelligence 29 (8) (2007) 1465–1469.
[26] C.-L. Liu, H. Sako, H. Fujisawa, Discriminative learning quadratic discriminant function for handwriting recognition, IEEE Trans. Neural Networks 15 (2) (2004) 430–444.
[27] C.-L. Liu, F. Yin, D.-H. Wang, Q.-F. Wang, CASIA online and offline Chinese handwriting databases, Proc. Int'l Conf. Document Analysis and Recognition, 2011, pp. 37–41.
[28] C.-L. Liu, F. Yin, D.-H. Wang, Q.-F. Wang, Online and offline handwritten Chinese character recognition: benchmarking on new databases, Pattern Recognition 46 (2013) 155–162.
[29] C.-L. Liu, F. Yin, Q.-F. Wang, D.-H. Wang, ICDAR 2011 Chinese handwriting recognition competition. Proc. Int'l Conf. Document Analysis and Recognition, 2011, pp. 1464–1469.
[30] H. Liu, X. Ding, Improve handwritten character recognition performance by heteroscedastic linear discriminant analysis, Proc. Int'l Conf. Pattern Recognition, 2006.
[31] M. Loog, R.P.W. Duin, Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion, IEEE Trans. Pattern Analysis and Machine Intelligence 26 (6) (2004) 732–739.
[32] M. Loog, R.P.W. Duin, R. Haeb-Umbach, Multiclass linear dimension reduction by weighted pairwise Fisher criteria, IEEE Trans. Pattern Analysis and Machine Intelligence 23 (7) (2001) 762–766.
[33] R. Lotlikar, R. Kothari, Fractional-step dimensionality reduction, IEEE Trans. Pattern Analysis and Machine Intelligence 22 (6) (2000) 623–627.
[34] C.R. Rao, The utilization of multiple measurements in problems of biological classification, J. Royal Statistical Society. Series B (Methodological) 10 (2) (1948) 159–203.
[35] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.
[36] B. Schölkopf, A. Smola, K.R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Computation 10 (5) (1998) 1299–1319.
[37] Y. Shao, C. Wang, B. Xiao, Fast self-generation voting for handwritten Chinese character recognition, Int'l J. Document Analysis and Recognition, 2012.
[38] A. Stuhlsatz, J. Lippel, T. Zielke, Feature extraction with deep neural networks by a generalized discriminant analysis, IEEE Trans. Neural Networks and Learning Systems 23 (4) (2012) 596–608.
[39] D. Tao, X. Li, X. Wu, S.J. Maybank, Geometric mean for subspace selection, IEEE Trans. Pattern Analysis and Machine Intelligence 31 (2) (2009) 260–274.
[40] J.B. Tenenbaum, V. De Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.
[41] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cognitive Neuroscience 3 (1) (1991) 71–86.
[42] Y.-W. Wang, X.-Q. Ding, C.-S. Liu, Neighbor class linear discriminant analysis, J. Pattern Recognition and Artificial Intelligence (in Chinese) 25 (3) (2012) 406–410.
[43] W.K. Wong, M. Sun, Deep learning regularized Fisher mappings, IEEE Trans. Neural Networks 22 (10) (2011) 1668–1675.
[44] B. Xu, K. Huang, C.-L. Liu, Dimensionality reduction by minimal distance maximization, Proc. Int'l Conf. Pattern Recognition, 2010, pp. 569–572.
[45] J. Yang, A.F. Frangi, J. Yang, D. Zhang, Z. Jin, KPCA plus LDA: a complete kernel Fisher discriminant framework for feature extraction and recognition, IEEE Trans. Pattern Analysis and Machine Intelligence 27 (2) (2005) 230–244.
[46] H. Yu, J. Yang, A direct LDA algorithm for high-dimensional data with application to face recognition, Pattern Recognition 34 (10) (2001) 2067–2070.
[47] Y. Yu, J. Jiang, L. Zhang, Distance metric learning by minimal distance maximization, Pattern Recognition 44 (3) (2011) 639–649.
[48] X.-Y. Zhang, C.-L. Liu, Confused distance maximization for large category dimensionality reduction, Proc. Int'l Conf. Frontiers in Handwriting Recognition, 2012, pp. 213–218.
[49] X.-Y. Zhang, C.-L. Liu, Writer adaptation with style transfer mapping, IEEE Trans. Pattern Analysis and Machine Intelligence (In Press), 2013.
[50] Y. Zhang, D.Y. Yeung, Worst-case linear discriminant analysis, Proc. Advances in Neural Information Processing Systems, 2010.
[51] M. Zhu, A.M. Martinez, Subclass discriminant analysis, IEEE Trans. Pattern Analysis and Machine Intelligence 28 (8) (2006) 1274–1286.

Xu-Yao Zhang received the B.S. degree in computational mathematics from Wuhan University, Wuhan, China, in 2008. He is currently pursuing a Ph.D. degree in pattern recognition and intelligent systems at the Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include pattern recognition, machine learning, and especially large category classification, dimensionality reduction and classifier adaptation.

Cheng-Lin Liu is a Professor at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation of Chinese Academy of Sciences, Beijing, China, and is now the deputy director of the laboratory. He received the B.S. degree in electronic engineering from Wuhan University, Wuhan, China, the M.E. degree in electronic engineering

from Beijing Polytechnic University, Beijing, China, the Ph.D. degree in pattern recognition and intelligent control from the Chinese Academy of Sciences, Beijing, China, in 1989, 1992 and 1995, respectively. He was a postdoctoral fellow at Korea Advanced Institute of Science and Technology (KAIST) and later at Tokyo University of Agriculture and Technology from March 1996 to March 1999. From 1999 to 2004, he was a research staff member and later a senior researcher at the Central Research Laboratory, Hitachi, Ltd., Tokyo, Japan. His research interests include pattern recognition, image processing, neural networks, machine learning, and especially the applications to character recognition and document analysis. He has published over 140 technical papers at prestigious international journals and conferences.