Neural Networks 34 (2012) 56–64

Contents lists available at SciVerse ScienceDirect

Neural Networks

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

Maxi-Min discriminant analysis via online learning

Bo Xu [a], Kaizhu Huang [b,*], Cheng-Lin Liu [b]

[a] Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun East Road Beijing 100190, PR China

[b] National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun East Road Beijing 100190, PR China

ARTICLE INFO

Article history:
Received 17 June 2011
Received in revised form 8 May 2012
Accepted 12 June 2012

Keywords:
Linear discriminant analysis
Dimensionality reduction
Multi-category classification
Handwritten Chinese character recognition

ABSTRACT

Linear Discriminant Analysis (LDA) is an important dimensionality reduction algorithm, but its performance is usually limited on multi-class data. Such limitation is incurred by the fact that LDA actually maximizes the average divergence among classes, whereby similar classes with smaller divergence tend to be merged in the subspace. To address this problem, we propose a novel dimensionality reduction method called Maxi-Min Discriminant Analysis (MMDA). In contrast to the traditional LDA, MMDA attempts to find a low-dimensional subspace by maximizing the minimal (worst-case) divergence among classes. This "minimal" setting overcomes the problem of LDA that tends to merge similar classes with smaller divergence when used for multi-class data. We formulate MMDA as a convex problem and further as a large-margin learning problem. One key contribution is that we design an efficient online learning algorithm to solve the involved problem, making the proposed method applicable to large scale data. Experimental results on various datasets demonstrate the efficiency and the efficacy of our proposed method against five other competitive approaches, and the scalability to the data with thousands of classes.

1. Introduction

Dimensionality reduction has been an important topic in machine learning and pattern recognition. Principal Component Analysis (PCA) (Gao, 2008) does not guarantee the discrimination performance as it does not consider the label information. Linear Discriminant Analysis (LDA), developed by Fisher in 1936, is a popular method that has achieved great success in many fields (Fukunaga, 1990). Under the homoscedastic Gaussian assumption, LDA is equivalent to finding the maximum-likelihood (ML) parameter estimates and leads to the optimal projection axis used for two-category data (Campbell, 2008). When applied to multi-category (e.g., c-category) data, LDA can still achieve good performance in many cases. Precisely speaking, Rao (1948) showed that $c-1$ dimensional subspace given by LDA, wherein c is the class number, and is also guaranteed to be Bayes optimal in multi-class homoscedastic case under the condition that the data features $d \geq c$. Fig. 1(a) is one example where LDA can find the good projection axis for three-class separation.

However, LDA may fail to find good projection for other multi-class data, especially when the category number is far larger than the data features (Loog, Duin, & Haeb-Umbach, 2001), e.g., in Chinese character recognition (with 3755 classes and merely several hundred features). In this case, it is impossible to reduce the dimensionality to any number equal to or slightly smaller than $c-1$. Fig. 1(b) illustrates a typical example that LDA fails to find a good projection when the dimensionality is reduced to 1 for a three-class problem. Clearly, by LDA, the transformed data of class 1 and class 2 would overlap with each other heavily, leading to worse performance for consequent classification. This problem of LDA, or more clearly, the phenomenon that LDA tends to merge similar or closer classes when the dimension of the projected subspace is strictly lower than the class number minus one, is called the class separation problem in the literature (Tao, Li, Wu, & Maybank, 2009). In contrast, the dashed axis in Fig. 1(b), would be a reasonable projection axis that can appropriately make the data of each class well separated.

The criterion of LDA is trying to search a low-dimensional subspace which can maximize the between-class covariance while minimizing the within-class covariance. Using Lemma 1 (provided in Section 3.2), LDA actually exploits an *average* setting, i.e., LDA tries to maximize the *average* divergences among different classes. The divergence of any two classes is defined as the distance between the mean vectors of the two classes in the whitening space. To maximize the *average* divergence, LDA tends to find the subspace preserving the larger divergences and ignoring the smaller divergences, as illustrated in Fig. 1(b). This causes the overlap of the similar classes, with smaller divergences, after data transformation.

To address the class separation problem of LDA, there have been several proposals in the literature. Loog et al. (2001)

* Corresponding author.
E-mail addresses: kzhuang@nlpr.ia.ac.cn, kaser.huang@gmail.com (K. Huang).

0893-6080/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2012.06.001

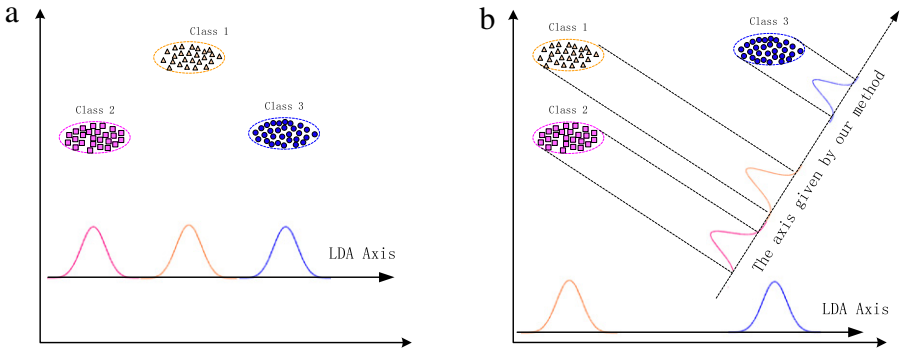


Fig. 1. Illustration of LDA for multi-class data.

developed a heuristic method called approximate Pairwise Accuracy Criterion (aPAC) that adds larger weights for similar classes in the estimation of the between-class covariance. Lotlikar and Kothari (2000) proposed the so-called Fractional-step Linear Discriminant Analysis (F-LDA) by heuristically and iteratively reducing dimension from a high-dimensional space to the low-dimensional space. Recently, Abou-Moustafa, de la Torre, and Ferrie (2010) designed the Pareto Discriminant Analysis (PDA) by forcing the pairwise distance to be equal after transformation. These methods usually deal with the class separation problem by imposing different weights on classes, either iteratively or directly. However, the weighting function is always ad-hoc and often needs to be adapted in different applications. For PDA, the involved optimization problem is non-convex, making its performance usually limited in practice. More related work can be referred to Section 2.

Unlike previous approaches, in this paper, a novel *worst-case* framework called Maxi-Min Discriminant Analysis (MMDA) is proposed. More specifically, instead of maximizing the *average* divergence among different classes, MMDA attempts to maximize the *minimal* (worst-case) divergence. In this worst-case setting, MMDA tries to push away each pair of classes with small divergence as large as possible. This consequently avoids the aforementioned problem and hence presents a more rigorous method (compared with aPAC and F-LDA). Obviously, the proposed MMDA method is still optimal for two-class problems under the homoscedastic Gaussian assumption, since it is degraded to the standard LDA when the class number is equal to two. Hence, the proposed worst-case method can be seen as a more generalized version of LDA for multi-class problems.

One important contribution of this paper is that we formulate the MMDA problem as a convex programming problem, or more precisely a Semi-Definite Programming (SDP) problem. Since SDP is computationally intractable even for medium-size data, we first transform the involved SDP problem to a large margin problem and then present an efficient online learning algorithm to solve it. The proposed online algorithm is important in that (a) it is computationally more efficient by removing the constraint of SDP, and (b) it has a nice convergence property. We note that Bian and Tao (2010) and Yu, Jiang, and Zhang (2011) proposed a similar model from the view point of distance metric learning or dimensionality reduction. However, their models are basically SDP problems and are hence intractable for large-scale data.

The paper is organized as follows. In Section 2, we detail the related work. In Section 3, we present our novel worst-case framework for dimensionality reduction in detail. The model definition, the theoretical justification and practical optimization will be discussed in turn in this section. In Section 4, we evaluate our algorithm and report experimental results. Finally, we set out the conclusion with some remarks.

A preliminary version of this paper has been early published in Xu, Huang, and Liu (2010), which is however significantly expanded both in theory and experiments in the current version.

2. Related work

There are a number of dimensionality reduction approaches related to our work (Abou-Moustafa et al., 2010; He & Niyogi, 2003; Loog et al., 2001; Lotlikar & Kothari, 2000; Sugiyama, 2006; Tang, Suganthan, Yao, & Qin, 2005; Tao et al., 2009; Yan et al., 2007).

Sugiyama (2006, 2007) developed the Local Fisher Discriminant Analysis (LFDA) method that combines the merits of Locality Preserving Projection (LPP) (He & Niyogi, 2003) into LDA. This method is shown to be very promising in many real datasets. However, it is mainly designed for solving classification tasks when classes distributions are multi-model and its performance is limited in handling the class separation problem. Loog et al. (2001) developed a method called approximate Pairwise Accuracy Criterion (aPAC) that adds larger weights for similar classes in the estimation of the between-class covariance. This method is well motivated and partially solves the class separation problem. However, it remains a problem how to select an optimal weighting function. Tang et al. (2005) proposed a relevance weighted LDA which incorporates the inter-class relationships as relevance weights into the estimation of the overall within-class scatter matrix in order to improve the performance of the basic LDA method. The major problem is still how to select the optimal weighting function. Another related approach called Fractional-step Linear Discriminant Analysis (F-LDA) is proposed in Lotlikar and Kothari (2000). F-LDA is a heuristic method, which can generate better classification accuracy by iteratively reducing dimension from a high-dimensional space to the low-dimensional space. This improves the robustness of choosing the weighting function. Its performance is limited in that a large number of steps should be used to collapse each dimension and the choice of a scaling parameter is always critical for the final result. Marginal Fisher Analysis (MFA) (Yan et al., 2007) is also highly related to our method. MFA characterizes the interclass compactness by the neighboring points in the same class and characterizes the interclass separability by the connection of marginal points. However, graph construction based on the whole data-set is time-consuming, which limits its application.

As a short summary, the above methods usually deal with the class separation problem by imposing different weights on classes, either iteratively or directly. However, the weighting function is always ad-hoc and often needs to be adapted in different applications. Recently, Pareto Discriminant Analysis was proposed to address the class-separation problem (Abou-Moustafa et al., 2010) by forcing the pairwise distance to be equal after transformation. There are two shortcomings for the method. On the one hand, it involves a non-convex programming problem; on the other hand, the so-called Pareto optimal criterion may essentially not be a good criterion because a bad local minimum can be a Pareto optimal point as well. Yu et al. (2011) proposed a similar criterion called minimal distance maximization for

distance metric learning. After being recognized as an instance of SDP, the problem is solved by many existing numerical packages. However, due to the large computational complexity of SDP, this model cannot be applied for large-scale datasets. Similarly, Bian and Tao (2010) designed a sequential SDP relaxation to solve the Maxi-Min criterion, which unfortunately has the large time cost due to the involvement of SDP.

We summarize the contributions of our paper as follows.

- We present a systematic and elegant approach to deal with the class separation problem. Maximization of the minimal divergence naturally and directly pushes away the closer classes and hence presents a more rigorous way to solve the class separation problem than the weighting methods. This is distinct with Loog et al. (2001), Lotlikar and Kothari (2000) and Tang et al. (2005).
- We formulate the model as a convex programming problem and further a large-margin learning problem. This is different from the heuristic methods or non-convex approaches, e.g., in Abou-Moustafa et al. (2010), Loog et al. (2001) and Lotlikar and Kothari (2000).
- We develop a novel online learning algorithm to solve the involved SDP problem. We also show that the designed online algorithm has a nice convergence property, which hence guarantees its learning performance. More importantly, the proposed online learning algorithm enables our method applicable to large-scale data, e.g., Chinese character data with over 1 million samples. This is different from the work Bian and Tao (2010) and Yu et al. (2011).

3. Framework of maxi-min discriminant analysis

In this section, we describe our novel dimensionality reduction framework MMDA. We first present the notation and a two-step view of LDA, and then introduce the detailed framework.

3.1. Notation

We first present the notation used in the paper. Let $x_i \in R^D (i = 1, 2, \ldots, n)$ be D-dimensional samples and $c_i \in \{1, 2, \ldots, c\}$ be their associated class labels, where n is the number of samples and c is the number of classes. Let n_i be the number of samples in the class i, $\sum_{i=1}^{l} n_i = n$. Let $X = [x_1, x_2, \ldots, x_n]$ represent all samples as a matrix. The purpose of linear dimensionality reduction is to find a projection matrix W which maps a D-dimensional data space to a d-dimensional subspace ($d < D$), $W^T : R^D \to R^d$, where $W = [w_1, w_2, \ldots, w_d]$. Let $Y = [y_1, y_2, \ldots, y_n]$, $y_i \in R^m$ represent all samples in the embedding space projected by matrix W^T, where y_i is given by $y_i = W^T x_i, i \in [1, n]$. Let m_i, m_i' and M_i be the mean of the class i in the original space, the whitening space, and the low-dimensional space respectively. In addition, $A \succeq 0$ means the matrix A is a positive semi-definite matrix and $\|A\|_F$ indicates the Frobenius norm of the matrix A.

3.2. Two-step view of LDA

Before we describe our novel MMDA framework, following Fukunaga (1990), we present a two-step view of LDA. The transformation matrix W of LDA is usually given by the eigenvalue decomposition of $S_W^{-1} S_B$, where S_W is the within-class covariance and S_B is the so-called between-class covariance. The solution of LDA can be equivalently computed in two steps by whitening S_W first and then conducting Principle Component Analysis (PCA) in the whitening space (Fukunaga, 1990).

Denote the eigenvectors of S_W by the matrix P and the eigenvalues by the matrix Λ_1, then we have $S_W = P \Lambda_1 P^T$. The first step of LDA is to transform S_W to an identity matrix using a whitening transformation matrix $W_1 = P \Lambda_1^{-1/2}$, i.e., $W_1^T S_W W_1 = I$. Accordingly, S_B is transformed to $W_1^T S_B W_1 = S_B'$. The second step of LDA is to utilize the PCA transformation matrix W' on class centers in the whitening space. This is equivalent to the transformation that maximizes the average divergence among all the classes in the whitening space. It is proved in Lemma 1. Thus the final transformation of LDA is the combination of the two steps, i.e., the whitening transformation W_1 and the PCA transformation W':

$$W = W_1 W' = P \Lambda_1^{-1/2} W'.$$

Lemma 1. *Under the homoscedastic Gaussian assumption, the LDA solution is a linear transformation maximizing the average divergence among all the classes in the whitening space, or in particular, LDA finds a projection matrix W' in the whitening space to satisfy the following criterion:*

$$\max_{W'} \left(\sum_{1 \le i < j \le l} \left\| M_i - M_j \right\|^2 \right). \tag{1}$$

Proof can be see in Tao et al. (2009).

Maximizing the average divergence will possibly lead to serious overlap of those similar classes (as illustrated in Fig. 1(b)). In the next subsection, we will show that our proposed worst-case framework MMDA can solve this problem.

3.3. The MMDA model

To attack the class separation problem, we present a novel method called MMDA that maximizes the minimal pairwise divergence among the class centers. In this worst-case setting, the closer classes will be pushed further away and this hence alleviates the overlap problem systematically.

Our approach, MMDA, also follows the two-step framework as LDA. After applying the whitening transformation on the data, it finds a projection matrix W' to satisfy the following criterion:

$$\max_{W'} \left(\min_{1 \le i < j \le l} \left\| \mathrm{M}_i - \mathrm{M}_j \right\|^2 \right). \tag{2}$$

The major difference from LDA is that, MMDA is maximizing the minimal divergence, while LDA is maximizing the average divergence (as proved in Lemma 1). Maximizing the average divergence may lead some classes to overlap in the transformed space, while maximizing the minimal divergence can avoid such cases effectively.

We now show how to solve the optimization problem of MMDA. Eq. (2) is equivalent to:

$$\max_{W', y} y \text{ s.t. } D_{ij} \ge y, \quad \forall i, j, 1 \le i < j \le l, \tag{3}$$

where $D_{ij} = \left\| \mathrm{M}_i - \mathrm{M}_j \right\|^2$ and $y > 0$. As D_{ij} can be rewritten as

$$\begin{aligned} D_{ij} &= tr\left(W'^T \left(m_i' - m_j' \right) \left(m_i' - m_j' \right)^T W' \right) \\ &= tr\left(W' W'^T \left(m_i' - m_j' \right) \left(m_i' - m_j' \right)^T \right), \end{aligned}$$

we can rewrite Eq. (3) as: $tr\left(A S_{ij}' \right) \ge y, \forall i, j, 1 \le i < j \le l$, where $A = W' W'^T$ and S_{ij}' is the scatter matrix between class i and class j in the whitening space.

After adding a reasonable constraint to matrix A for avoiding trivial solution, i.e., $\|A\|_F = 1$, our model can be rewritten as:

$$\max_{A} y \text{ s.t. } \begin{cases} tr\left(A S_{ij}' \right) \ge y, & \forall i, j, 1 \le i < j \le l \\ \|A\|_F = 1 \\ A \succeq 0. \end{cases} \tag{4}$$

The problem of (4) forms a typical Semi-Definite Programming (SDP) problem, which can be solved by some software packages, e.g. Sedumi (Sturm, 1999) or CSDP (Borchers, 1999).

Once we obtain matrix A, the next problem is how to compute W' from A. Here we apply the least square method to solve W':

$$J\left(W'\right) = \min_{W'} \left\|A - W'W'^{T}\right\|^2. \tag{5}$$

It is easy to prove that the optimal W' is the first d largest eigenvectors of A.

Therefore, the final transformation of MMDA is the combination of the whitening transformation W_1 and W':

$$W = W_1 W'. \tag{6}$$

Remarks. To find the transformation matrix, the MMDA method is needed to solve a typical SDP problem. This is similar to the work in Bian and Tao (2010) and Yu et al. (2011). However, SDP is computationally intractable for high dimensional and large-scale data. To address this challenge, we propose an efficient online learning algorithm which avoids the SDP problem but with a nice convergence property. We will introduce the online algorithm in the next subsection.

3.4. The MMDA online algorithm

In this subsection, we firstly transform the problem (4) into a large margin optimization problem, and then relax it to a soft-margin problem. After that, we discuss an efficient online algorithm to solve the problem.

Lemma 2. *The problem (4) is equivalent to the following large margin optimization problem:*

$$\min_{Q} \frac{1}{2}\|Q\|_F^2$$
$$s.t. \begin{cases} tr(QS'_{ij}) \geq 1, & \forall i, j, 1 \leq i < j \leq l \\ Q \succeq 0. \end{cases} \tag{7}$$

In the above, Q is proportional to the matrix A in (4), subject to a constant factor, i.e., $Q = kA$, where k is a positive value.

Proof. We can rewrite $A = \frac{A'}{\|A'\|_F}$ and remove $\|A\|_F = 1$. (4) is equivalent to the following optimization problem:

$$\max_{A'} y^2$$
$$\text{s.t.} \begin{cases} tr\left(\frac{A'}{\|A'\|_F} S'_{ij}\right) \geq y, & \forall i, j, 1 \leq i < j \leq l \\ A' \succeq 0. \end{cases} \tag{8}$$

Then, we have

$$\max_{A'} y^2$$
$$\text{s.t.} \begin{cases} tr\left(\frac{A'}{\|A'\|_F \cdot y} S'_{ij}\right) \geq 1, & \forall i, j, 1 \leq i < j \leq l \\ A' \succeq 0. \end{cases} \tag{9}$$

By defining $\frac{A'}{\|A'\|_F y} = Q$, then we have

$$\min_{Q} \frac{1}{2}\|Q\|_F^2$$
$$\text{s.t.} \begin{cases} tr\left(QS'_{ij}\right) \geq 1, & \forall i, j, 1 \leq i < j \leq l \\ Q \succeq 0. \end{cases} \tag{10}$$

This completes the proof. □

Similar to other large margin methods, we can also introduce a soft-margin problem:

$$\min_{Q} \frac{\lambda}{2}\|Q\|_F^2 + \sum_{ij} \xi_{ij}$$
$$\text{s.t.} \begin{cases} tr\left(QS'_{ij}\right) \geq 1 - \xi_{ij}, & \forall i, j, 1 \leq i < j \leq l \\ Q \succeq 0 \\ \xi_{ij} \geq 0, & \forall i, j, 1 \leq i < j \leq l. \end{cases} \tag{11}$$

Replacing the objective in Eq. (11) with $Dm_t = \{m'_i, m'_j\}$ yields:

$$f\left(Q; Dm_t\right) = \frac{\lambda}{2}\|Q\|_F^2 + \xi_{ij}, \tag{12}$$

where $\xi_{ij} = \min\{0, 1 - tr(QS'_{ij})\}$.

The sub-gradient of the above objective can be given by:

$$\nabla_{f_Q} = \lambda Q - \ell\left[tr\left(QS'_{ij}\right) < 1\right] S'_{ij}, \tag{13}$$

where $\ell\left[tr\left(QS'_{ij}\right) < 1\right]$ is the indicator function which takes a value of one if its arguments is true and zero otherwise. We then update $Q_{t+1} \leftarrow Q_t - \eta_t \nabla_t$ using a step size of $\eta_t = 1/(\lambda t)$:

$$Q_{t+1} \leftarrow \left(1 - \frac{1}{t}\right) Q_t + \eta_t \ell\left[tr\left(QS'_{ij}\right) < 1\right] S'_{ij}. \tag{14}$$

The final output Q_{t+1} is achieved after a predetermined number of iteration. We can write the solving method in Algorithm 1.

Algorithm 1 Online learning algorithm for Maxi-Min Discriminant Analysis

```
INPUT: Predefined learning rate λ, training samples
Initialize Q_0 = 0
for t = 1, ..., T do
    Receive a pair of training examples m_i, m_j. (m'_i, m'_j are center of class i and class j after whitening.)
    if (m'_i − m'_j)^T Q_t (m'_i − m'_j) < 1 then
        Q_{t+1} = (1 − 1/t) Q_t + η_t S'_ij. (η_t = 1/(λt))
    else
        Q_{t+1} = (1 − 1/t) Q_t.
    end if
end for
```

Note that compared to the SDP solution, the proposed online learning algorithm is advantageous in that (i) it is computationally more efficient by avoiding solving an SDP problem and (ii) it has a proved bound on the average instantaneous objective.

3.5. Convergence analysis

In this subsection, we analyze the convergence property of the MMDAOnline algorithm. First, it is easy to verify that the proposed algorithm converges when T goes to infinity, since Q_t will stay unchanged if T is sufficiently large. In the following, we will then present an analysis to bound the average instantaneous objective of MMDAOnline algorithm such that this average instantaneous objective value given by the online algorithm will be tightly bounded by the optimal objective value. After that, we provide probabilistic analysis to show that the final solution terminated at epoch t (t is sufficiently large) will usually lead to good performance.

We first borrow a lemma from Hazan, Kalai, Kale, and Agarwal (2006) and Shalev-Shwartz, Singer, and Srebro (2007).

Lemma 3. *Let $f_1, \ldots, f_T$ be a sequence of λ-strongly convex functions. Let B be a closed convex set and define $\Pi_B(\omega) = \arg\min_{\omega' \in B} \|\omega - \omega'\|$. Let $\omega_1, \ldots, \omega_{T+1}$ be a sequence of vectors such that $\omega_1 \in B$ and for $t \geq 1$, $\omega_{t+1} = \Pi_B(\omega_t - \eta_t \nabla_t)$, where ∇_t belongs to the sub-gradient set of f_t at ω_t and $\eta_t = 1/(\lambda t)$. Assume that for all t, $\|\nabla_t\| \leq G$. Then, for all $u \in B$ we have*

$$\frac{1}{T}\sum_{t=1}^{T} f_t(w_t) \leq \frac{1}{T}\sum_{t=1}^{T} f_t(u) + \frac{G^2(1+\ln(T))}{2\lambda T}. \tag{15}$$

In the above, if $f(w) - \lambda/2\|w\|^2$ is a convex function, then the function f is called λ-strong convex. Based on Lemma 3, we are now ready to provide the convergence property of the MMDAOnline algorithm in Theorem 1.

Theorem 1. *Assume the difference of any pair of means in the whitening space is at most R, i.e., $\forall i \neq j$, $\|m_i' - m_j'\| \leq R$. Let $f(Q)$ be given by the original objective function of (11) and $f(Q; Dm_t)$ be defined by Eq. (12). Let $Q^* = \arg\min_Q f(Q)$ and let $c = 4R^4$. Then for $T \geq 3$,*

$$\frac{1}{T}\sum_{t=1}^{T} f(Q_t; Dm_t) \leq \frac{1}{T}\sum_{t=1}^{T} f\left(Q^*; Dm_t\right) + \frac{c(1+\ln(T))}{2\lambda T}. \tag{16}$$

Proof. The updating algorithm can be rewritten as:

$$Q_{t+1} = Q_t - \eta_t \nabla_t, \tag{17}$$

where ∇_t is defined in Eq. (13).

Based on the Lemma 3, Eq. (16) is established if we prove that the following conditions hold:

1. $f(Q; Dm_t)$ is λ-strongly convex.
2. $\|\nabla_t\| \leq 2R^2$.

Proof of condition (1): It is clear that $f(Q; Dm_t)$ is a λ-strongly convex, since $f(Q; Dm_t) = \frac{\lambda}{2}\|Q\|_F^2 + \xi_{ij}$ is λ-strongly convex function plus a convex function.

Proof of condition (2): The updating step can be rewritten as

$$Q_{t+1} = \left(1 - \frac{1}{t}\right) Q_t - \frac{1}{t\lambda} v_t, \tag{18}$$

where $v_t = \ell\left[tr\left(QS_{ij}'\right) < 1\right] S_{ij}'$. Therefore, the initial weight of each v_i is $1/\lambda i$ and then on rounds $j = i+1, \ldots, t$, it will be multiplied by $1 - 1/j$. Thus the overall weight of v_i in Q_{t+1} is

$$\frac{1}{\lambda i}\prod_{j=i+1}^{t}\frac{j-1}{j} = \frac{1}{\lambda t} \tag{19}$$

which implies that we can rewrite Q_{t+1} as

$$Q_{t+1} = \frac{1}{\lambda t}\sum_{i=1}^{t} v_i. \tag{20}$$

From the above, we immediately have that $\|Q_{t+1}\| \leq R^2/\lambda$ and $\|S_{ij}'\| = \|m_i' - m_j'\|^2 \leq R^2$. Recalling that $\nabla_t = \lambda Q_t - \ell\left[tr\left(Q_t S_{ij}'\right) < 1\right] S_{ij}'$, we can therefore get $\|\nabla_t\| \leq 2R^2$. This completes the proof. □

We start to obtain a bound on $f(Q_t)$, which is evaluated at a single Q_t. We borrow a lemma from Kakade and Tewari (2009).

Lemma 4 (*Corollary 7 in Kakade and Tewari (2009)*)**.** *Assume that the conditions stated in Theorem 1 hold and that each pair of means are sampled uniformly at random. Assume also that $R \geq 1$ and $\lambda \leq 1/4$. Then, with a probability of at least $1 - \delta$ we have*

$$\frac{1}{T}\sum_{t=1}^{T} f(Q_t) - f\left(Q^*\right) \leq \frac{21c\ln(T/\delta)}{\lambda T}. \tag{21}$$

Furthermore, due to the convexity of f, we have the following inequality:

$$f\left(\frac{1}{T}\sum_{t=1}^{T} Q_t\right) \leq \frac{1}{T}\sum_{t=1}^{T} f(Q_t). \tag{22}$$

Using the above inequality and Lemma 4, we can immediately obtain Corollary 1.

Corollary 1. *Assume that the conditions stated in Lemma 4 hold and let $\bar{Q} = \frac{1}{T}\sum_{t=1}^{T} Q_t$. Then, with probability of at least $1 - \delta$ we have*

$$f\left(\bar{Q}\right) \leq f\left(Q^*\right) + \frac{21c\ln(T/\delta)}{\lambda T}. \tag{23}$$

Corollary 1 states that the average of Q_t can reach a sufficiently good solution with a high probability for the involved optimization problem, provided that T could be sufficiently large. In practice, the final solution Q_{T+1} obtained at the epoch T can often lead to better performance. To show this, we now prove that by at least half chance, Q_{T+1} is good.

Lemma 5. *Assume that the conditions stated in Lemma 4 hold. Then, if t is selected at random from $[T]$, we have with a probability of at least $\frac{1}{2}$ that*

$$f(Q_t) \leq f\left(Q^*\right) + \frac{42c\ln(T/\delta)}{\lambda T}. \tag{24}$$

Proof. Define a random variable $Z = f(Q_t) - f(Q^*)$ where the randomness is over the choice of the index t. From the definition of Q^* as the minimizer of $f(Q)$ we clearly have that Z is a non-negative random variable. Thus, from Markov inequality $P[Z \geq 2E[Z]] \leq \frac{1}{2}$, the claim now follows by combining the fact that $E[Z] = \frac{1}{T}\sum_{t=1}^{T} f(Q_t) - f(Q^*)$ with the bound given in Lemma 4. This completes the proof. □

From the above lemma, we know that the last hypothesis Q_{t+1} achieves an accurate solution in at least half of the cases by a random termination during iteration. Therefore, it is reasonable to evaluate the error of the last hypothesis at a random stopping time. The above lemma tells us that we are likely to obtain a good solution after two attempts on average.

4. Experimental results

We conduct extensive experiments to verify both the efficiency and the efficacy of the proposed algorithm for dimensionality reduction. We compare our algorithm to the following five state-of-the-art algorithms, PCA, LDA, LPP (He & Niyogi, 2003), LFDA (Sugiyama, 2006) and aPAC (Loog et al., 2001) on nine datasets (including two illustrative datasets, five UCI datasets, and two benchmark large-scale character datasets). Particularly, we evaluate our model on a large-scale Chinese character data with over 1 million samples and 3755 classes. We call the MMDA solved by SDP as MMDASDP, and the one solved by the online algorithm as MMDAOnline. We set the maximum number of iterations for our MMDAOnline to be 1000, and obtain the learning rate λ by cross validation. All the algorithms are implemented and run using Matlab on a PC with 3.0 GHz CPU and 2G RAM.

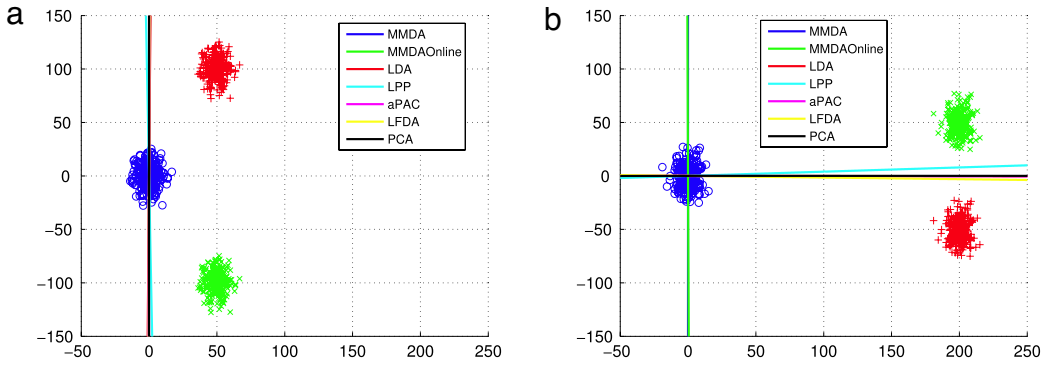


Fig. 2. Projection directions of different methods. In (a), all the methods show similar performance; in (b), our proposed MMDA can separate the classes correctly (the projection axes given by MMDASDP and MMDAOnline coincide with each other), while the comparison approaches (generating horizontal projection axes) tend to merge the closer classes, i.e., the green and red data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

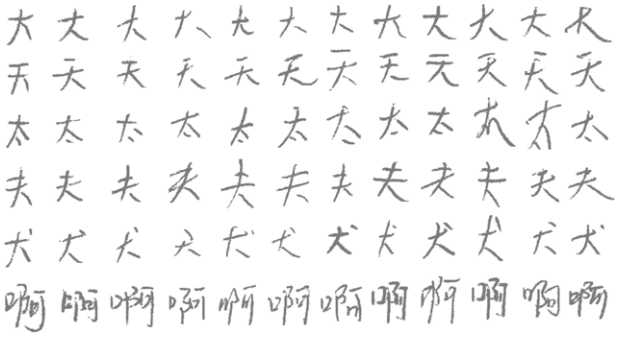


Fig. 3. Images of similar Chinese characters. Each row shows 12 samples for each character. The first 5 rows look very similar, while the last row is significantly different.

4.1. Results on synthetic data

To validate the effectiveness of the proposed MMDA, we first conduct experiments on a synthetic dataset. We generated 250 samples for each of the three classes (750 samples in total). Moreover, the samples in each class were obtained from a two-dimensional standard normal distribution. Each approach finds a projection axis to separate different classes as much as possible. The projection axes are shown in Fig. 2. In Fig. 2(a), all the approaches successfully separate different classes: their projection axes are all overlapped. However, when applied to (b) with similar classes, only MMDA can separate classes correctly in the Fig. 2(b) (the projection axes given by MMDASDP and MMDAOnline coincide with each other). The other methods generate horizontal projection axes and tend to merge closer classes, i.e., the green and red data. This clearly shows the superiority of MMDA over the other approaches for multi-class problems.

4.2. Results on similar Chinese character data

In this experiment, in order to further illustrate the effectiveness of our proposed MMDA model, we select six characters '大天太夫犬啊' from the benchmark Handwritten Chinese character dataset CASIA-HWDB1.1 (details about the dataset can be seen shortly in Section 4.5). As observed from these six characters, the first 5 look very similar, while the remaining one is very different. These 6 classes have 1495 samples in total. Fig. 3 provides some samples in this dataset.

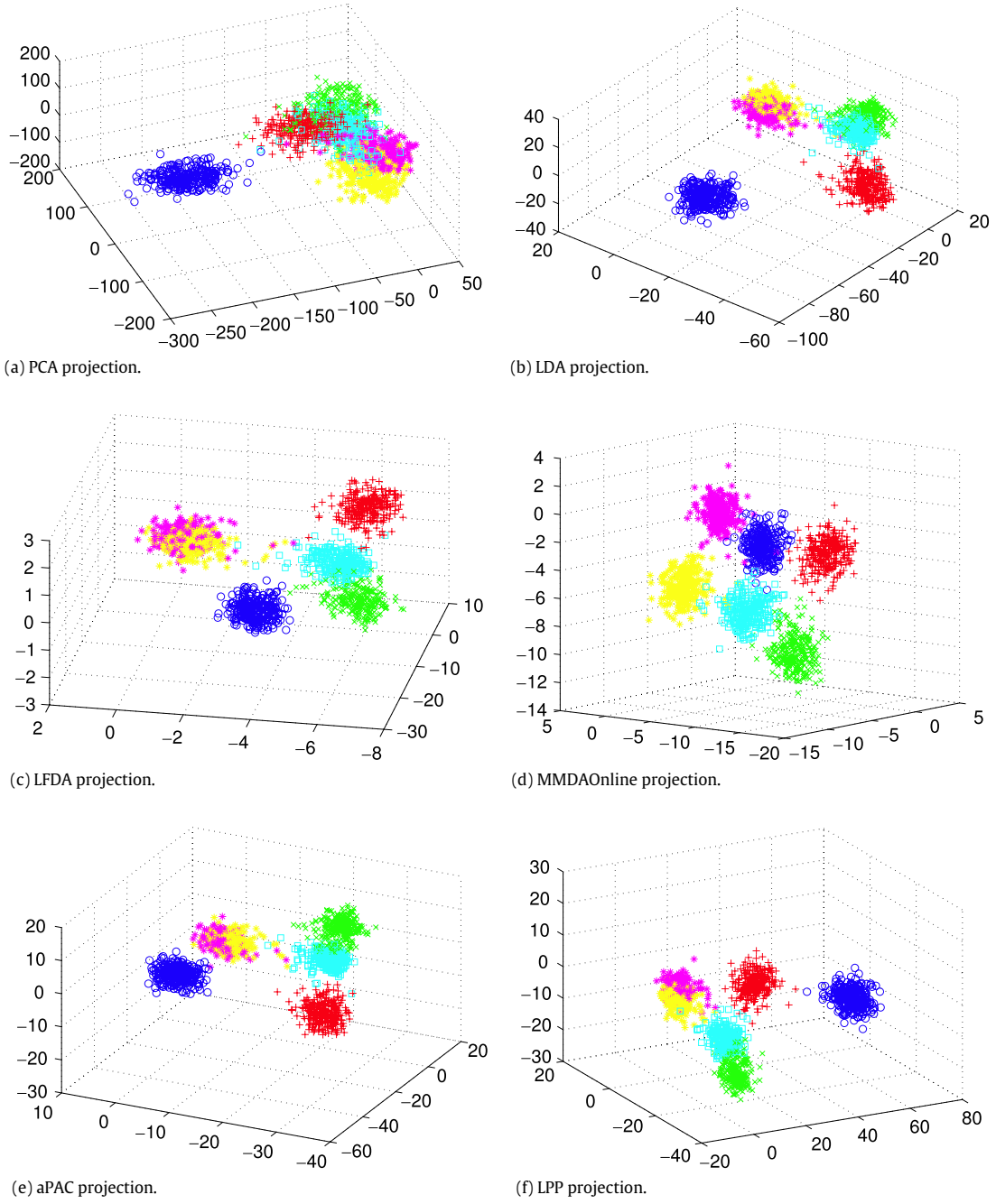
The image samples were firstly normalized by the Line Density Interpolation (LDI) (Liu & Marukawa, 2005) algorithm, then the normalization-cooperated gradient feature (NCGF) (Liu, 2007) were extracted. The gradient elements were decomposed into 8 directions and each direction was extracted 8×8 values by Gaussian blurring. This leads to the final feature dimensionality as 512.

To better illustrate the effectiveness of different algorithms, we apply each approach to reduce the dimensionality of the dataset to 3. We then visualize the samples in Fig. 4. Clearly observed, PCA generally cannot discriminate different classes, hence most similar classes overlap with each other. LDA also tends to merge similar characters due to its "average" setting. In comparison, our proposed MMDA algorithm adopts a worst-case setting which can better separate different classes. In addition, other algorithms, e.g., LFDA, aPAC, LPP, can generate better performance than PCA, but still mix several classes seriously.

4.3. Results on UCI data

After illustrating the effectiveness of our proposed model, we conduct extensive experiments of the data classification on the following five datasets from the UCI repository (Asuncion & Newman, 2007): (1) Teaching, with 3 classes, 5 features and 151 instances; (2) Wine, with 3 classes, 13 features and 178 instances; (3) Balance-scale, with 3 classes, 4 features and 625 instances; (4) Sat-log, with 6 classes, 36 features, 4435 training instances and 2000 test instances; (5) optdigits, with 10 classes, 64 features, 3823 training instances and 1797 test instances. For simplicity, we specify the reduced dimensionality as the $c-1$ in the experiments on UCI data. After the dimensionality reduction, the k Nearest Neighbor (k-NN) is then adopted as the classifier to evaluate the performance of each approach. We compute the recognition rate of 1, 3, 5, 7, 9 nearest neighbor, and record the best rate classifier. The reported test accuracies are acquired using 10-fold Cross Validation (CV) for the first 3 small- or medium-size datasets. For Sat-log and optdigits, the accuracies are calculated on their specified test sets.

The recognition results are reported in Table 1. Clearly, MMDA (including the batch model and the online version) demonstrates overall best performance than the other methods, especially when compared to LDA and PCA. aPAC also demonstrates good performance. As discussed before, aPAC is also well justified for solving the class separation problem. However, it needs to define weighting functions beforehand, which is somehow ad-hoc. We also perform the statistical test on Teaching, Wine and Balance. The t-test results under the 5% significant level show the superiority of our method. More particularly, although MMDA is just marginally better in Teaching, it is significantly better than the other comparison algorithms in Wine. Moreover, in Balance,



(a) PCA projection.

(b) LDA projection.

(c) LFDA projection.

(d) MMDAOnline projection.

(e) aPAC projection.

(f) LPP projection.

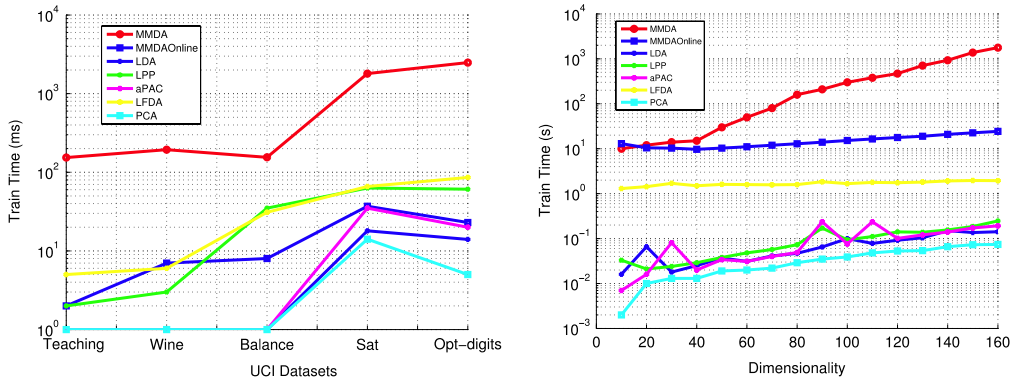
Fig. 4. Illustration of different algorithms. The proposed MMDA algorithm demonstrates the best separation ability.

Table 1
Classification rate on UCI data (mean ± std-dev%).

Datasets	MMDA	MMDAOnline	LDA	LPP	aPAC	LFDA	PCA
Teaching	96.6 ± 1.6	**97.5 ± 0.2**	96.5 ± 1.6	96.4 ± 1.8	96.6 ± 1.8	96.1 ± 1.9	96.4 ± 1.8
Wine	96.4 ± 1.5	**96.7 ± 0.9**	92.9 ± 1.9	87.3 ± 2.6	95.6 ± 1.7	95.8 ± 1.0	77.9 ± 3.2
Balance	98.3 ± 0.6	**98.5 ± 0.4**	97.7 ± 0.8	97.8 ± 0.5	97.8 ± 0.6	98.3 ± 0.7	72.1 ± 4.4
Sat	87.1	87.9	88.4	88.3	88.5	87.5	**88.9**
Opt-digits	96.0	**96.5**	95.7	95.4	95.0	95.8	95.3

MMDA shows performance similar to that of LFDA but it is significantly better than the remaining algorithms.

To examine the learning efficiency of the proposed online algorithm, we also plot the curves of its computational time of each model, especially MMDAOnline against MMDASDP on different datasets. The result is shown in Fig. 7(a). It is evident that the proposed online algorithm overcomes the problem incurred by SDP and significantly reduces the training time compared with its batch version. It demonstrates comparable learning efficiency against the other methods (see Fig. 5).



(a) UCI dataset. (b) USPS dataset.

Fig. 5. Comparison on training time.

Table 2
Classification rate on USPS data.

Dimensionality	MMDAOnline	LDA	LPP	LFDA	aPAC	PCA
9	**89.8**	89.7	89.4	89.2	89.7	88.2
8	**89.6**	89.0	88.9	88.7	88.3	87.4
7	**89.5**	88.3	88.2	86.8	88.3	86.0
6	87.5	87.2	87.0	85.4	**88.2**	82.1
5	**86.4**	84.1	84.5	84.0	86.3	78.6
4	**83.6**	78.3	78.7	78.4	82.2	75.6
3	**73.5**	63.9	64.2	66.6	72.9	61.6

4.4. Results on USPS data

In this section, we report the experimental results of the proposed algorithm using a well-known large-scale character recognition dataset, the United States Postal Services (USPS) dataset, in which there are 9298 handwriting character measurements divided into 10 classes. The entire USPS dataset is divided into two parts, a training set with 7291 measurements and a test set with 2007 measurements. Each measurement is a vector with 256 dimensions.

We apply different algorithms to the USPS dataset and report the k-NN's accuracy when the reduced dimensionality is set from 3 to 9. As MMDASDP is generally an SDP problem and hence cannot be applied on this large-scale data, we do not report its accuracy. Table 2 shows the final results. As observed, MMDAOnline demonstrates the best overall performance against other comparison methods. When compared with LDA, its performance is significantly better when the dimensionality is equal to 5, 4, 3. This again shows the advantages of the proposed method.

Meanwhile, to examine the learning efficiency of the proposed online algorithm against SDP more carefully, we intentionally train the MMDA on a different number of features, which are gradually increased from 10 features to 160 features (by random selection). The reduced dimensionality is fixed to 9 for simplicity. We record the training time for each experiment. The results are plotted in Fig. 7(b). As observed clearly, the training time of the online algorithm remains almost unchanged, while the SDP solving method MMDASDP is already very slow even when only 140 features are kept for training. When over 200 features are used, MMDASDP crashes due to the out-of-the-memory problem. The online algorithm presents one of the major contributions for this paper.

4.5. Results on large-scale Chinese character data

In this section, we further examine the proposed MMDAOnline algorithm on the large-scale Chinese character dataset CASIA-HWDB1.1 (Liu, Yin, Wang, & Wang, 2011). CASIA-HWDB1.1 is a new dataset of unconstrained Chinese handwritten characters collected by National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences (CASIA). The handwritten data were generated using Anoto pen on paper such that both online and off-line data can be obtained. We use the off-line isolated handwritten characters dataset CASIA-HWDB1.1 as our evaluation data. CASIA-HWDB1H1 contains over 1 million samples including a training set with 897,758 training samples and a test set with 223,991 test samples. There are 3755 classes in the dataset. Fig. 6 illustrates some samples from the CASIA-HWDB1H1 dataset.

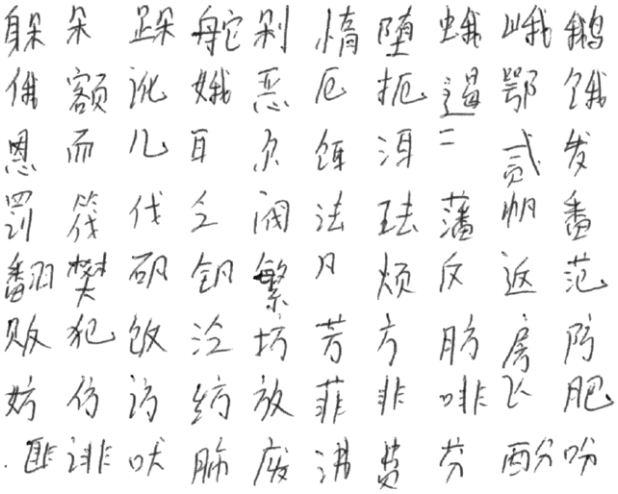
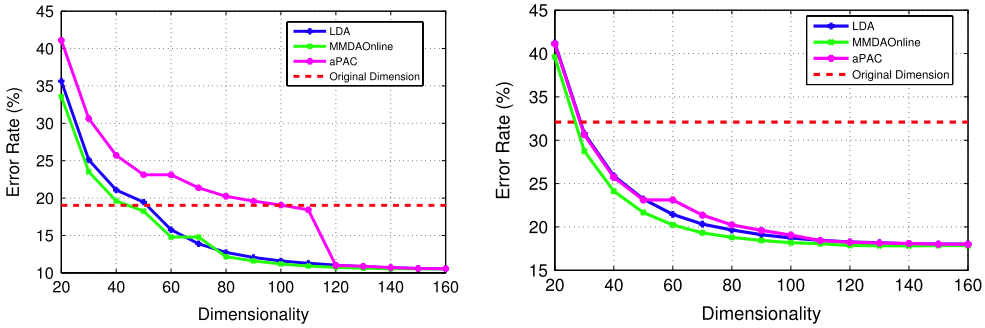


Fig. 6. Samples of CASIA-HWDB1H1.

We use the same pre-processing method as mentioned in Section 4.2 on the dataset. This finally leads to a 512-dimensional feature for each sample. We apply different dimensionality reduction algorithms to reduce the features of each sample to a specified number. The transformed samples were then fed into the classifier to measure the performance of different dimensionality reduction algorithms. As k-NN is too slow for the large-scale data, we adopt the popular Nearest Class Mean (NCM) (Fukunaga, 1990) and the Modified Quadratic Discriminant Analysis (MQDF) (Kimura, Takashina, Tsuruoka, & Miyake, 1987; Xu, Huang, Zhu, King, & Lyu, 2009) as the classifier. These two classifiers are generally the state-of-the-art classifiers in Chinese character recognition. For simplicity, we compare our proposed algorithm merely with the two traditional algorithms, LDA and aPAC on this database, as the other algorithms are widely recognized not suitable for Chinese character classification in the literature.

We specify the reduced dimensionality from 160 to 20 gradually and then report the classification performance accordingly.



(a) MQDF classifier. (b) Nearest center classifier.

Fig. 7. Error rate on CASIA-HWDB1H1 dataset.

Fig. 7(a) shows the error rate with the MQDF classifier and Fig. 7(b) reports the error rate with the NCM classifier. From the experiment results, we can see that our algorithm performs consistently better than LDA and aPAC for all the dimensionality. The difference is even more distinct as the dimensionality is small. This clearly shows the advantages of our proposed algorithm.

Moreover, to verify the effectiveness of dimensionality reduction, we also report the error rates of the MQDF and NCM classifiers by the red dash line on the original dimensionality (without dimensionality reduction) in Fig. 7(a) and (b) respectively. Obviously, after dimensionality reduction, both the Nearest-center classifier and MQDF classifier can achieve a lower error rate than that in the original data space.

5. Conclusion

LDA is one of the most important subspace methods in machine learning research and applications; however, for a c-class classification task, it tends to merge together nearby classes if the dimension of the projected subspace is strictly lower than $c - 1$. To address this problem, we proposed a novel criterion, MMDA, based on the maximization of the minimal divergence among the different classes. This solves the class separation problem of LDA. More importantly, we presented an efficient MMDA online algorithm, making our model applicable on very large-scale data. Extensive experiments on seven datasets and two additional large-scale datasets demonstrated the effectiveness and efficiency of the MMDA algorithm over the five state-of-the-art comparison methods.

There are many issues deserving our attentions. First, we mainly address the linear transforms for dimensionality reduction. The extension of our approach to its non-linear version is one of our focuses in the near future. Second, although we develop an online algorithm which significantly speeds up the learning efficiency, it still remains very interesting if our algorithm can be further sped up, e.g., by implementing in the parallel style. Finally, it remains an open problem regarding how to choose a suitable dimensionality for reduction. In practice, we may have to use cross validation in order to search for a good dimensionality, which is however computationally expensive. We will leave this topic for future work.

Acknowledgments

This work was supported by National Basic Research Program of China (973 Program) Grants 2012CB316301 and 2012CB316302, National Natural Science Foundation of China (NSFC) Grants No. 61075052, No. 60825301, the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant XDA06030300) and Tsinghua National Laboratory for Information Science and Technology (TNList) Cross-discipline Foundation.

References

Abou-Moustafa, K.T., de la Torre, F., & Ferrie, F.P. 2010. Pareto discriminant analysis. In: *Proceedings of the computer vision and pattern recognition*, CVPR. pp. 3602–3609.
Asuncion, A., & Newman, D. 2007. UCI machine learning repository.
Bian, W., & Tao, D. (2010). Max–min distance analysis by using sequential SDP relaxation for dimension reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(5), 1037–1050.
Borchers, B. (1999). CSDP, AC library for semidefinite programming. *Optimization Methods and Software*, *11*(1), 613–623.
Campbell, N. A. (2008). Canonical variate analysis — a general model formulation. *Australian & New Zealand Journal of Statistics*, *26*(1), 86–96.
Fukunaga, K. (1990). *Introduction to statistical pattern recognition*. Academic Press.
Gao, J. (2008). Robust L1 principal component analysis and its Bayesian variational inference. *Neural Computation*, *20*(2), 555–572.
Hazan, E., Kalai, A., Kale, S., & Agarwal, A. (2006). Logarithmic regret algorithms for online convex optimization. *Learning Theory*, 499–513.
He, X., & Niyogi, P. (2003). Locality preserving projections. *Advances in neural information processing systems (NIPS)*, 153–160.
Kakade, S., & Tewari, A. (2009). On the generalization ability of online strongly convex programming algorithms. *Advances in Neural Information Processing Systems*, *21*, 801–808.
Kimura, F., Takashina, K., Tsuruoka, S., & Miyake, Y. (1987). Modified quadratic discriminant functions and the application to Chinese character recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *9*(1), 149–153.
Liu, C.-L. (2007). Normalization-cooperated gradient feature extraction for handwritten character recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(8), 1465–1469.
Liu, C.-L., & Marukawa, K. (2005). Pseudo two-dimensional shape normalization methods for handwritten Chinese character recognition. *Pattern Recognition*, *38*(12), 2242–2255.
Liu, C.-L., Yin, F., Wang, D., & Wang, Q. (2011). Casia online and offline Chinese handwriting databases. In *Proceedings of the 11th international conference on document analysis and recognition*, ICDAR. (pp. 37–41). IEEE.
Loog, M., Duin, R. P. W., & Haeb-Umbach, R. (2001). Multiclass linear dimension reduction by weighted pairwise Fisher criteria. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(7), 762–766.
Lotlikar, R., & Kothari, R. (2000). Fractional-step dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(6), 623–627.
Rao, C. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, *10*(2), 159–203.
Shalev-Shwartz, S., Singer, Y., & Srebro, N. (2007). Pegasos: primal estimated sub-gradient solver for svm. In *Proceedings of the 24th international conference on machine learning* (pp. 807–814). ACM.
Sturm, J. F. (1999). Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, *11*, 625–653.
Sugiyama, M. (2006). Local Fisher discriminant analysis for supervised dimensionality reduction. In *Proceedings of the 23rd international conference on machine learning (ICML)* (pp. 905–912). ACM.
Sugiyama, M. (2007). Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *The Journal of Machine Learning Research*, *8*, 1027–1061.
Tang, E. K., Suganthan, P. N., Yao, X., & Qin, A. K. (2005). Linear dimensionality reduction using relevance weighted LDA. *Pattern Recognition*, *38*(4), 485–493.
Tao, D., Li, X., Wu, X., & Maybank, S. J. (2009). Geometric mean for subspace selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(2), 260–274.
Xu, B., Huang, K., & Liu, C.-L. 2010. Dimensionality reduction by minimal distance maximization. In: *Proceedings of the 20th international conference on pattern recognition*, ICPR. pp. 569–572.
Xu, Z., Huang, K., Zhu, J., King, I., & Lyu, M. (2009). A novel kernel-based maximum a posteriori classification method. *Neural Networks*, *22*(7), 977–987.
Yan, S., Xu, D., Zhang, B., Zhang, H., Yang, Q., & Lin, S. (2007). Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(1), 40–51.
Yu, Y., Jiang, J., & Zhang, L. (2011). Distance metric learning by minimal distance maximization. *Pattern Recognition*, *44*(3), 639–649.