Contents lists available at ScienceDirect

Pattern Recognition

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

Generalized re-weighting local sampling mean discriminant analysis

Jing Chai *, Hongwei Liu, Zheng Bao

National Laboratory of Radar Signal Processing, Xidian University, Xi'an, Shanxi, China

ARTICLE INFO

Article history:
Received 21 November 2009
Received in revised form
8 April 2010
Accepted 19 April 2010

Keywords:
Pattern recognition
Linear discriminant analysis
Local sampling mean discriminant analysis
Generalized re-weighting framework
Classification

ABSTRACT

Despite the general success in the pattern recognition community, linear discriminant analysis (LDA) has four intrinsic drawbacks. In this paper, we propose a new feature extraction algorithm, namely, local sampling mean discriminant analysis (LSMDA), to make up for the first three drawbacks, and a generalized re-weighting (GRW) framework to make up for the fourth drawback. Extensive experiments are conducted on both synthetic and real-world datasets to evaluate the classification performance of our work. The experimental results demonstrate the effectiveness of both LSMDA and the GRW framework in classifications.

1. Introduction

Feature extraction plays an important role in many machine learning fields, e.g., unsupervised clustering, semi-supervised learning, supervised classification, etc. [1,2]. Among various feature extraction techniques, subspace based approaches, which seek linear or nonlinear transformations that project data from the original space to low-dimensional representative subspaces, take a dominant position and have attracted general attention in the last decades. Principal component analysis (PCA) [3] and linear discriminant analysis (LDA) [4], as well as their kernelized extensions [5,6], are two representative and perhaps most widely utilized subspace feature extraction algorithms. PCA, also termed as the Karhunen–Loeve transformation, which seeks a linear transformation by minimizing the reconstruction error, is a typical unsupervised feature extraction technique and usually adopted as a preprocessing tool to reduce noises. Different from PCA, LDA makes use of the class label information and aims at maximizing the ratio of between-class scatter to within-class scatter in the transformed subspace. In general, due to its supervised learning nature, LDA possesses stronger discriminative ability than unsupervised techniques (such as PCA), and is hence widely utilized in various pattern recognition (PR) fields.

Despite the general success in the PR community, LDA has several intrinsic drawbacks, of which the first one is that it requires homogeneous samples to be Gaussian distributed. In many real-world applications, this requirement is too strict to satisfy, and hence, LDA may perform poorly. To cope with this problem, many researchers propose to extract data's local structural information, instead of the global one (what LDA extracts), to construct scattering matrices. Typical algorithms of this kind include nonparametric discriminant analysis (NDA) [7–9], marginal Fisher analysis (MFA) [10], local discriminant embedding (LDE) [11], subclass discriminant analysis (SDA) [12], etc. The main difference between these algorithms and LDA lies in that the former utilize neighboring sampling pairs or pairs consisting of samples and subclass sampling means (to calculate subclass sampling means, we should first perform clustering on samples in the same class by some clustering algorithm, e.g., the nearest neighbor (NN) clustering algorithm [12], and then calculate the sampling means for each cluster and get corresponding subclass sampling means), in contrast with the latter's pairs consisting of samples and sampling means, to describe data's scattering status. In some sense, these algorithms can be treated as local relaxations of LDA, as they assume homogeneous samples are locally Gaussian distributed, i.e., samples within a small neighborhood area, or a subclass, are Gaussian distributed.

The second drawback of LDA is that the number of available projection vectors is up-bounded with $\min\{D,c-1\}$, where D and c, respectively, denote data's initial dimensionality and the number of classes. Hence, for applications with very few classes, the number of available projection vectors will be very small, and hence the resulting subspace may be unable to provide sufficient discriminative information for classifications. Aforementioned local algorithms, e.g., NDA, MFA, LDE, and SDA, can alleviate this limitation, since their replacement of sampling means by neighbors or subclass sampling means may increase the number of independent terms in scattering matrices (especially in the

* Corresponding author. Tel.: +86 29 88206441; fax: +86 29 88201448.
E-mail address: jingchai@yahoo.cn (J. Chai).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.04.017

between-class scattering matrix). Another technique, namely, optimal discriminant projection pursuit (ODPP) [13], alleviates this limitation in a different way. Unlike algorithms that attempt to solve a generalized Rayleigh quotient problem (such as LDA, NDA, MFA, LDE, SDA, etc.), ODPP first constructs a candidate projection set, and then searches over this set to pick up those most discriminative projection vectors (through an AdaBoost [14] process), and thus utilizes the picked projection vectors to construct the ultimate transformation matrix. Because the AdaBoost searching process is quite time-consuming, the time complexity of ODPP is usually very high, and this may limit applications of ODPP, especially for cases where efficiency is a crucial evaluating criterion.

In order to weaken the disadvantages caused by the aforementioned two drawbacks, we propose a new model, namely, local sampling mean, based on which the newly constructed scattering matrices not only no longer require homogeneous samples to be Gaussian distributed, but also increase the upper bound of available projection dimensions to min{D,n} (n denotes the number of total samples). Different from other local learning models such as neighbors and subclass sampling means, local sampling means utilize more potential information (information from all samples in a given class, not from neighboring samples or samples in a given subclass), which may provide additional help for enhancing discriminability.

The third drawback is that all discrepant vectors (a discrepant vector denotes the difference of two vectors, e.g., $x_1 = x_2 - x_3$) are treated equivalently during the construction of scattering matrices. In fact, different discrepant vectors contribute differently to classifications, and the equivalent treatment to all discrepant vectors is not a good choice. Alternatively, we propose a weighting scheme, through which discrepant vectors are endowed with different weighting coefficients according to their corresponding samples' classifiability, and hence, expect to construct scattering matrices more appropriately.

In the LDA framework, there exists numerous generalized eigenvectors corresponding to one generalized eigenvalue, and all these generalized eigenvectors are optimal with respect to the Fisher criterion [15]. At first sight, it seems that the norms of LDA projection vectors[1] do not contribute to classifications, and many researchers neglect the influence of norms. This negligence is thus the fourth drawback of LDA. Ma et al. [15] demonstrated that the norms of projection vectors do have strong influence on classification performance and proposed a re-weighting (RW) approach to learn the optimal norm of projection vectors for a nearest mean (NM) classifier. By adding a shrinking objective term and extending the RW approach to be applicable for any subspace based feature extraction algorithm and any prototype based classifier, we propose a generalized re-weighting (GRW) framework, with which better generalization abilities are expected to be obtained.

In short, we list the main contributions of this paper as follows:

(1) A new feature extraction algorithm, namely, local sampling mean discriminant analysis (LSMDA), is proposed to make up for the first three drawbacks of traditional LDA and to cope with more general classification tasks.
(2) A generalized re-weighting (GRW) framework is proposed to make up for the fourth drawback of LDA. More importantly, the proposed framework is applicable for any subspace based feature extraction algorithm and any prototype based classifier, and this openness makes it possible for researchers to design new relevant algorithms.
(3) The connection (similarities and dissimilarities) of our work to some related work is analyzed, which gives a further understanding on both our and related work and may provide some guidance to select appropriate algorithms for various applications.

The rest of this paper is organized as follows. In Section 2, a short description of LDA is reviewed. In Section 3, the design of LSMDA is discussed in detail. The GRW framework is introduced in Section 4. In Section 5, the relationship of our work to some related work is analyzed. Experimental comparisons of our work with several competing algorithms are given in Section 6. Finally, we give concluding remarks and descriptions of future work in Section 7.

2. Brief review of LDA

In this section, we give a short description of LDA and start with an introduction of some useful notations. Suppose the training set is composed of c classes {1,…,c}, and $x_j^i \in \mathbb{R}^D$ is a D-dimensional column vector denoting the jth sample in class i; n_i and $n = \sum_{i=1}^{c} n_i$, respectively, denote the number of samples in class i and that in all classes, while $m_i = (1/n_i)\sum_{j=1}^{n_i} x_j^i$ and $m = (1/n)\sum_{i=1}^{c}\sum_{j=1}^{n_i} x_j^i$, respectively, denote the sampling mean for class i and that for all classes.

On the basis of above definitions, LDA can be expressed as

$$G_{\mathrm{LDA}} = \arg\max_{G} \frac{\mathrm{trace}(G^{\mathrm{T}} S_{\mathrm{LDA}}^{\mathrm{b}} G)}{\mathrm{trace}(G^{\mathrm{T}} S_{\mathrm{LDA}}^{\mathrm{w}} G)} \tag{1}$$

where G_{LDA} denotes the required transformation matrix, $S_{\mathrm{LDA}}^{\mathrm{w}}$ and $S_{\mathrm{LDA}}^{\mathrm{b}}$, respectively, denote the within-class and between-class scattering matrices:

$$S_{\mathrm{LDA}}^{\mathrm{w}} = \sum_{i=1}^{c}\sum_{j=1}^{n_i} (x_j^i - m_i)(x_j^i - m_i)^{\mathrm{T}} \tag{2}$$

$$S_{\mathrm{LDA}}^{\mathrm{b}} = \sum_{i=1}^{c} n_i (m_i - m)(m_i - m)^{\mathrm{T}} \tag{3}$$

Eq. (1) can be solved analytically by operating generalized eigenvalue decomposition:

$$S_{\mathrm{LDA}}^{\mathrm{b}} u = \lambda S_{\mathrm{LDA}}^{\mathrm{w}} u \tag{4}$$

Suppose that $\lambda_1 > \cdots > \lambda_d$ are the first d largest generalized eigenvalues and $u_1, \ldots, u_d$ are the corresponding generalized eigenvectors, the solution of Eq. (1) can be expressed as

$$G_{\mathrm{LDA}} = [u_1, \ldots, u_d] \tag{5}$$

3. Design of LSMDA

In this section, we introduce a new feature extraction algorithm, namely, local sampling mean discriminant analysis (LSMDA), to make up for the first three drawbacks of LDA. We start our discussion with a new definition: local sampling mean.

Definition 1. Local sampling mean

Suppose that the affinity matrix $A \in \mathbb{R}^{n \times n}$ is calculated as

$$A(x_p^i, x_q^j) = \exp\left(-\frac{\|x_p^i - x_q^j\|_2^2}{\sigma_1}\right)$$

[1] Here the norm of projection vectors is something different from the exact definition of norm in mathematics; in fact it deals with each projection vector as a feature and denotes the weighting coefficients of these features; here we adopt the term 'norm' as per Ref. [14], which first took this problem into account.

where σ_1 is the kernel width parameter. Given a query sample $x_p{}^i$, its local sampling mean for class j is defined as

$$m_{(x_p^i,j)}=\frac{\sum\limits_{q=1}^{n_j}A(x_p^i,x_q^j)x_q^j}{\sum\limits_{q=1}^{n_j}A(x_p^i,x_q^j)} \tag{6}$$

Moreover, if $j=i$, $m_{(i,j)}$ is called $x_p{}^i$'s homogeneous local sampling mean; otherwise it is a heterogeneous one.

Different from sampling mean, local sampling mean is 'query-dependent', i.e., its calculation is dependent on the affinity between samples in some class and the query sample, with each sample's weighting coefficient degrading exponentially with increase of squared Euclidean distance between it and the query sample. Thereby, large coefficients are endowed on samples close to the query sample and small ones on samples far away from the query sample. In other words, more attention is paid to the contribution of local information, which explains the term 'local sampling mean'. The parameter σ_1 controls the varying speed of weighting coefficients with respect to distances.

The introduction of local sampling mean is based on the following two considerations. Firstly, local sampling mean is a weighted average of all samples in some class, and it utilizes more potential information than other local models such as neighbors and subclass sampling means; owing to this the construction of latter two models just utilizes partial information (within neighborhood area or a subclass) in some class. Secondly, note that if σ_1 is large enough, local sampling mean becomes sampling mean; thus sampling mean can be seen as a special case of local sampling mean. Hence, for applications where traditional LDA performs satisfactorily and for which introduction of local information cannot help improve or may even degrade the classification performance, we can expect local sampling means to reduce to sampling means.

On the basis of the local sampling mean model, within-class and between-class scattering matrices, as well as the corresponding transformation matrix, can be redefined as

$$S^{\mathrm{w}}_{\mathrm{new}}=\sum_{i=1}^{c}\sum_{p=1}^{n_i}(x_p^i-m_{(x_p^i,i)})(x_p^i-m_{(x_p^i,i)})^{\mathrm{T}} \tag{7}$$

$$S^{\mathrm{b}}_{\mathrm{new}}=\sum_{i=1}^{c}\sum_{p=1}^{n_i}\sum_{j\neq i,j=1}^{c}(x_p^i-m_{(x_p^i,j)})(x_p^i-m_{(x_p^i,j)})^{\mathrm{T}} \tag{8}$$

$$G_{\mathrm{new}}=\arg\max_G\frac{\mathrm{trace}(G^{\mathrm{T}}S^{\mathrm{b}}_{\mathrm{new}}G)}{\mathrm{trace}(G^{\mathrm{T}}S^{\mathrm{w}}_{\mathrm{new}}G)} \tag{9}$$

By replacing Eqs. (2) and (3) with Eqs. (7) and (8), the disadvantages caused by the first and second drawbacks of traditional LDA can be alleviated. Next we give a synthetic example with multimodal distributed data to show how the transformation of Eq. (9) makes up for the first drawback of LDA intuitively. The synthetic data consist of two classes, with each datum containing two dimensions. For class 1, data are drawn from two Gaussian distributions with equal prior probability, i.e., $p_{11}=p_{12}$ and $\mu_{11}=\{5,5\}$, $\mu_{12}=\{-5,-5\}$, $\Sigma_{11}=\Sigma_{12}=2I$; for class 2, data are drawn from another two Gaussian distributions with equal prior probability, i.e., $p_{21}=p_{22}$ and $\mu_{21}=\{4,-4\}$, $\mu_{22}=\{-4,4\}$, $\Sigma_{21}=\Sigma_{22}=I$. We generate 200 samples for each class and show the distribution of these samples in Fig. 1. Moreover, we also plot the projection direction of LDA (Line 1) and that of Eq. (9) (Line 2). It is noted that by projecting samples onto Line 2, two classes can be well separated, while by projecting samples onto Line 1, two classes mix together and it is difficult to separate them reasonably. Through Fig. 1, the performance improvement by replacing sampling means with local sampling means is shown clearly. Generally speaking, for multimodal distributed data, sampling mean is not a favorable first-order statistic to describe data's distributing status. In particular, for the example shown in Fig. 1, by minimizing the within-class scatter, samples are forced to move towards their homogeneous sampling means, which leads to heavy mixtures due to the overlapping of two sampling means. Comparatively, different local sampling means can be well separated, and, hence, samples in different classes are prevented from mixing together by adopting the local sampling mean model.

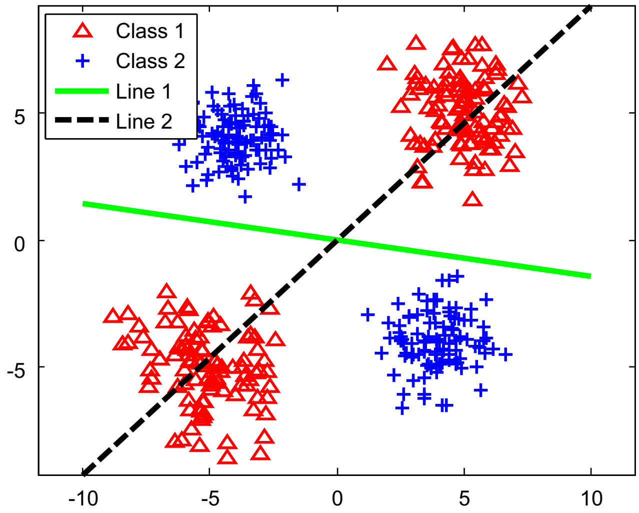

Fig. 1. Illustration of multimodal distributed data. The local sampling means are calculated with $\sigma_1=4$.

By replacing sampling means with local sampling means, the upper bound of available projection dimensions comes to min{D,n}, which is usually larger than traditional LDA's min{D,$c-1$}, especially for cases with very few classes. Through the above replacement, we may have more available projection vectors, which may provide us with more discriminative information for classifications.

The third drawback of LDA is that all discrepant vectors are treated equivalently, i.e., they attract the same attention during the construction of scattering matrices. However, this equivalent treatment may not be a good choice. Given two samples, suppose one can be easily distinguished from samples in other classes and the other is difficult to distinguish; then intuitively, an appropriate choice (in the sense of classification) might be that discrepant vectors corresponding to the difficult-to-classify sample become more important than those corresponding to the easy-to-classify one. By integrating the introduction of the local sampling mean model and the consideration on different samples' classifiability, we propose local sampling mean discriminant analysis (LSMDA), a new feature extraction algorithm that seeks the following transformation:

$$G_{\mathrm{LSMDA}}=\arg\max_G\frac{\mathrm{trace}(G^{\mathrm{T}}S^{\mathrm{b}}_{\mathrm{LSMDA}}G)}{\mathrm{trace}(G^{\mathrm{T}}S^{\mathrm{w}}_{\mathrm{LSMDA}}G)} \tag{10}$$

where the within-class and between-class scattering matrices, $S^{\mathrm{w}}_{\mathrm{LSMDA}}$ and $S^{\mathrm{b}}_{\mathrm{LSMDA}}$ are, respectively, defined as

$$S^{\mathrm{w}}_{\mathrm{LSMDA}}=\sum_{i=1}^{c}\sum_{p=1}^{n_i}w^{\mathrm{w}}_{(x_p^i,i)}(x_p^i-m_{(x_p^i,i)})(x_p^i-m_{(x_p^i,i)})^{\mathrm{T}} \tag{11}$$

$$S^{\mathrm{b}}_{\mathrm{LSMDA}}=\sum_{i=1}^{c}\sum_{p=1}^{n_i}\sum_{j\neq i,j=1}^{c}w^{\mathrm{b}}_{(x_p^i,j)}(x_p^i-m_{(x_p^i,j)})(x_p^i-m_{(x_p^i,j)})^{\mathrm{T}} \tag{12}$$

where $w^{\mathrm{b}}_{(x_p^i,j)}$ and $w^{\mathrm{w}}_{(x_p^i,i)}$ are, respectively, the weighting coefficient of the between-class discrepant vector $x_p^i - m_{(x_p^i,j)}$ and that of the within-class discrepant vector $x_p^i - m_{(x_p^i,i)}$:

$$w^{\mathrm{b}}_{(x_p^i,j)} = \exp\left(\frac{1-\eta_{ijp}}{1+\eta_{ijp}}/\sigma_2\right) \tag{13}$$

$$w^{\mathrm{w}}_{(x_p^i,i)} = \frac{1}{c-1}\sum_{j \neq i, j=1}^{c} w^{\mathrm{b}}_{(x_p^i,j)} \tag{14}$$

where

$$\eta_{ijp} = \|x_p^i - m_{(x_p^i,j)}\|_2 / \|x_p^i - m_{(x_p^i,i)}\|_2 \tag{15}$$

Note that η_{ijp} is a nonnegative parameter that reflects ${x_p}^i$'s classifiability. When $\eta_{ijp} > 1$, ${x_p}^i$ has high classifiability, tends to be correctly classified, and $w^{\mathrm{b}}_{(x_p^i,j)}$ lies in $(0,1)$; when $0 \leq \eta_{ijp} < 1$, ${x_p}^i$ has low classifiability, tends to be wrongly classified to belong to class j, and $w^{\mathrm{b}}_{(x_p^i,j)} > 1$; when $\eta_{ijp}=1$, the critical status arises, and $w^{\mathrm{b}}_{(x_p^i,j)} = 1$. Hence, through the properties of η_{ijp}, $w^{\mathrm{b}}_{(x_p^i,j)}$ also reflects ${x_p}^i$'s classifiability (large value depicts low classifiability and vice versa) and can be adopted to weight $x_p^i - m_{(x_p^i,j)}$ in the between-class scattering matrix. The parameter σ_2 controls the varying speed of $w^{\mathrm{b}}_{(x_p^i,j)}$ with respect to η_{ijp}, e.g., the smaller the σ_2, the steeper the variation.

The weighting coefficient $w^{\mathrm{w}}_{(x_p^i,i)}$ controls the relative importance $x_p^i - m_{(x_p^i,i)}$ occupies in the within-class scattering matrix and is simply the average of $w^{\mathrm{b}}_{(x_p^i,j)}$ for all $j \neq i$. Thereby, $w^{\mathrm{w}}_{(x_p^i,i)}$ denotes the average classifiability of ${x_p}^i$. When ${x_p}^i$ has low average classifiability, $w^{\mathrm{w}}_{(x_p^i,i)}$ is large; then $x_p^i - m_{(x_p^i,i)}$ will get a large weighting coefficient in the within-class scattering matrix. As a result, in the transformed subspace, difficult-to-classify samples will be forced strongly to move towards their homogeneous local sampling means, and this movement implicitly improves the separation of different classes.

4. Generalized re-weighting (GRW) framework

The Fisher optimization criterion is invariant to the scale of projection vectors, i.e., if u_i is a generalized eigenvector of Eq. (4), so is $\alpha_i u_i$ for any scalar α_i. However, they are not all optimal in the sense of classification. In [15], a re-weighting approach is proposed to learn the optimal norm of LDA projection vectors by minimizing the ranking loss for an NM classifier. According to this approach, learning the optimal norm of LDA projection vectors is equivalent to seeking the transformation

$$\tilde{G}_{\mathrm{LDA}} = [\alpha_1 u_1, \ldots, \alpha_d u_d],$$

where $u_1,\ldots,u_d$ are the same as those in Eq. (5), and we need to estimate only the coefficients $\alpha_1,\ldots,\alpha_d$. Let $\beta_l = {\alpha_l}^2$. Then the squared Euclidean distance between samples x_1 and x_2 in the LDA transformed subspace can be expressed as

$$\begin{aligned}\|\tilde{G}_{\mathrm{LDA}}^{\mathrm{T}} x_1 - \tilde{G}_{\mathrm{LDA}}^{\mathrm{T}} x_2\|_2^2 &= (x_1-x_2)^{\mathrm{T}} \tilde{G}_{\mathrm{LDA}} \tilde{G}_{\mathrm{LDA}}^{\mathrm{T}} (x_1-x_2) \\ &= \sum_{l=1}^{d} \alpha_l^2 [(x_1-x_2)^{\mathrm{T}} u_l]^2 = \sum_{l=1}^{d} \beta_l [(x_1-x_2)^{\mathrm{T}} u_l]^2\end{aligned} \tag{16}$$

The re-weighting (RW) approach in [15] is expressed as

$$\begin{aligned}&\min_{\gamma,\beta} \sum_{i=1}^{c}\sum_{p=1}^{n_i} \gamma_{ip} + C\sum_{l=1}^{d} \beta_l \\ &\text{s.t.} \quad \forall i,j,p: i,j \in \{1,\ldots,c\},\ i \neq j, p \in \{1,\ldots,n_i\} \\ &\qquad \|\tilde{G}_{\mathrm{LDA}}^{\mathrm{T}} x_p^i - \tilde{G}_{\mathrm{LDA}}^{\mathrm{T}} m_i\|_2^2 - \|\tilde{G}_{\mathrm{LDA}}^{\mathrm{T}} x_p^i - \tilde{G}_{\mathrm{LDA}}^{\mathrm{T}} m_j\|_2^2 \leq \xi_{ijp} - 1 \\ &\qquad \gamma_{ip} = \max_j \xi_{ijp} \\ &\quad \forall l: l \in \{1,\ldots,d\} \\ &\qquad \beta_l \geq 0\end{aligned} \tag{17}$$

The first constraint term is introduced to impose soft margins between different classes. In the transformed subspace, the squared Euclidean distance between each sample and its homogeneous sampling mean is expected to be smaller than that between it and the heterogeneous one, and without loss of generality, the difference is expected to be larger than 1. This unit difference, combined with the slack variables ξ_{ijp}, is introduced to constitute soft classification margins; γ_{ip} is a nonnegative parameter that denotes the maximum of ξ_{ijp}; when $\gamma_{ip}=0$, ${x_p}^i$ is correctly classified and does not violate the marginal constraint; when $0 < \gamma_{ip} < 1$, ${x_p}^i$ is correctly classified but violates the classification margins; when $\gamma_{ip} \geq 1$, ${x_p}^i$ tends to be misclassified. Hence, the first objective term reflects the total misclassification of the training set, and minimizing this term controls the training errors. The second objective term is the squared l_2-norm of the coefficient vector $\alpha = [\alpha_1,\ldots,\alpha_d]^{\mathrm{T}}$, and the aim in minimizing this term is to get the solution that utilizes as few features as possible. Due to $\beta_l = {\alpha_l}^2$, we restrict $\beta_l \geq 0$ naturally. Furthermore, we introduce a nonnegative parameter C to control the trade-off between two objective terms.

By replacing $\tilde{G}_{\mathrm{LDA}}$ with $\tilde{G}_{\mathrm{NDA}}$, Ma et al. [15] considered the problem of learning the optimal norm of NDA projection vectors for an NM classifier as well. In practice, except for the combination of LDA/NDA, and NM classifiers, we can naturally extend the re-weighting approach for any subspace based feature extraction algorithm and any prototype based classifier. Moreover, by adding a shrinking objective term, we get the following generalized re-weighting (GRW) framework:

$$\begin{aligned}&\min_{\gamma,\beta} \sum_{i=1}^{c}\sum_{p=1}^{n_i} \gamma_{ip} + C\sum_{l=1}^{d} \beta_l + C_e \|\tilde{G}_{\mathrm{SUB}}^{\mathrm{T}} x_p^i - \tilde{G}_{\mathrm{SUB}}^{\mathrm{T}} v_{(x_p^i,i)}\|_2^2 \\ &\quad \text{s.t.} \quad \forall i,j,p: i,j \in \{1,\ldots,c\},\ i \neq j, p \in \{1,\ldots,n_i\} \\ &\qquad \|\tilde{G}_{\mathrm{SUB}}^{\mathrm{T}} x_p^i - \tilde{G}_{\mathrm{SUB}}^{\mathrm{T}} v_{(x_p^i,i)}\|_2^2 - \|\tilde{G}_{\mathrm{SUB}}^{\mathrm{T}} x_p^i - \tilde{G}_{\mathrm{SUB}}^{\mathrm{T}} v_{(x_p^i,j)}\|_2^2 \leq \xi_{ijp} - 1 \\ &\qquad \gamma_{ip} = \max_j \xi_{ijp} \\ &\quad \forall l: l \in \{1,\ldots,d\} \\ &\qquad \beta_l \geq 0\end{aligned} \tag{18}$$

where $G_{\mathrm{SUB}} = [u_1,\ldots,u_d]$ and $\tilde{G}_{\mathrm{SUB}} = [\alpha_1 u_1, \ldots, \alpha_d u_d]$, respectively, denote the initial transformation of some given subspace based feature extraction algorithm and the required transformation that considers the problem of learning the optimal norm of projection vectors; $v_{(x_p^i,i)}$ and $v_{(x_p^i,j)}$ $(i \neq j)$, respectively, denote ${x_p}^i$'s homogeneous prototype and its heterogeneous prototype for class j.

Note that there is an additional objective term

$$C_e \|\tilde{G}_{\mathrm{SUB}}^{\mathrm{T}} x_p^i - \tilde{G}_{\mathrm{SUB}}^{\mathrm{T}} v_{(x_p^i,i)}\|_2^2$$

where C_e is a nonnegative parameter that controls the relative importance of this term with respect to the other two terms. This term forces samples to move towards their homogeneous prototypes. Intuitively, if all samples shrink within small areas around their homogeneous prototypes, samples in different classes can be well distinguished, as long as their homogeneous prototypes do not mix with each other. Through the introduction of this term, we expect to further distinguish samples in different classes, and, hence, improve the resulting classification accuracies. Moreover, this term resembles the sum-of-squared-error cost in K-means clustering [16], with the difference that each class is treated as a cluster and sampling means are replaced by given prototypes.

The GRW framework learns a new Mahalanobis distance metric and in fact serves as a post-processing procedure of the former dimensionality reduction part. In the dimensionality reduction part, the feature extraction algorithm learns a possibly slant low dimensional subspace, but the axis of each dimension is of unit length. This unit-length restriction is very strict, since it highly restricts subspaces' scale degrees of freedom and is usually not optimal in the sense of classifications. By endowing more degrees of freedom on the scale of subspaces, the GRW framework breaks this unit-length restriction and serves as a scale-adjustment procedure. Most importantly, through the processing of the GRW framework, it is possible to learn the optimal scale of former-learned subspaces for given classifiers.

Except for the nonlinear quasi-convex constraint $\gamma_{ip}=\max_j \xi_{ijp}$, all objective terms and constraints of Eq. (18) are linear and convex. Note that γ_{ip} is in fact the l_∞-norm of the vector $\xi_{ip}=[\xi_{i1p},\ldots,\xi_{i(i-1)p},\ \xi_{i(i+1)p},\ldots,\xi_{icp}]^{\mathrm{T}}$, i.e., $\gamma_{ip}=\|\xi_{ip}\|_\infty$. According to [17], we can approximate the nonlinear l_∞-norm constraint with the following linear constraints:

$$\gamma_{ip}\geq\xi_{ijp}\quad\forall i\neq j \tag{19}$$

Through this approximation, all constraints are linear and we can cope up with our GRW framework as a linear programming (LP) problem. In practice, usually a large number of training samples can be well distinguished, and ξ_{ip} is a sparse vector (most slack variables ξ_{ijp} are zeros). Thus, their pairwise distances will not incur the hinge loss, and we obtain very few active constraints (which incur the hinge loss). Similar to [15], we utilize active constraints only to solve Eq. (18), which can improve computational efficiency obviously.

By permuting and recombining various subspace based feature extraction algorithms and various classification prototypes, we may obtain many GRW realizations, all of which can be integrated into the GRW framework of Eq. (18). Table 1 shows several representative examples in this framework, by combining five feature extraction algorithms: LDA, NDA, MFA, SDA, and LSMDA, as well as three classification prototypes: nearest neighbor (NN), sampling mean (SM), and local sampling mean (LSM). For example, if $\tilde{G}_{\mathrm{SUB}}$ denotes the desired transformation of LDA, i.e., $\tilde{G}_{\mathrm{SUB}}=\tilde{G}_{\mathrm{LDA}}$, and $v_{(x_p^i,i)}$ and $v_{(x_p^i,j)}$, respectively, denote $x_p{}^{i}$'s nearest neighbor in class i and class j, then Eq. (18) comes under GRW–LDA–NN. Other examples in this framework can be deduced by analogy.

5. Related work

In this section, the connection of our work to some related work, including NDA [7–9], LMNN (large margin nearest neighbor classification) [18], LESS (lowest error in a sparse subspace) [19], and LMFW (large margin feature weighting) [20], is discussed and we expect to give further understandings on both our work and these works.

Based on the new definition of nonparametric between-class scattering matrix, NDA was proposed to get out of control of the intrinsic parametric nature of traditional LDA. Similar to LSMDA, NDA also intends to make up for the first three drawbacks of LDA. However, NDA realizes this intention incompletely, since its difference from LDA just lies in the between-class scattering matrix, while its within-class scattering matrix is still the same as that of LDA. Comparatively, LSMDA redefines both two scattering matrices and expects to obtain further performance improvements.

ODPP lies outside the Fisher framework and is quite different from LSMDA and other variations of LDA. Similar to the GRW framework, ODPP considers the relative importance of different projection vectors as well; however, there are two main differences between them: first, if we treat each projection vector as a feature, then the GRW framework can be seen as a feature weighting of these projections, while ODPP just performs a feature selection; second, ODPP is solved by AdaBoost with local optimal solutions, while the GRW framework is solved by linear programming, which can obtain global optimal solutions owing to its convex property.

LMNN, LESS, and LMFW are three large margin related approaches, either for feature weighting or feature extraction. The GRW framework is closely related to these approaches, since the first objective term of our framework, associated with the first constraint term, also constitutes large classification margins by learning the optimal norm of projection vectors. Note that learning the norm of projection vectors is equivalent to performing feature weighting in the transformed subspace. Thus, a main difference between the GRW framework and the above three approaches is that the former operates feature weighting on transformed samples, while the latter operate feature weighting or feature extraction on initial samples (samples without any transformation). Furthermore, LMNN, LESS, and LMFW operate either feature weighting or feature extraction; however, our work deals with both of these two aspects, i.e., our ultimate intention is to seek a transformation that corresponds to feature extraction, while in the GRW learning step, feature weighting is performed to seek the optimal norm of projection vectors. In short, the GRW framework refers to both aspects and gives an integrated procedure for learning subspace based feature extraction algorithms.

6. Experiments

We conduct extensive experiments on synthetic dataset, benchmark datasets, and radar HRRP dataset, to test the performances of the LSMDA algorithm and the GRW framework. Performance comparison is operated on the following competing approaches: LDA, NDA, MFA, SDA, LSMDA, and their GRW learning for an NN classifier, i.e., GRW–LDA–NN, GRW–NDA–NN, GRW–MFA–NN, GRW–SDA–NN, GRW–LSMDA–NN, as well as ODPP. An NN classifier is adopted as the classification tool to test the recognition performances of all these approaches. For ODPP, the numbers of candidate projection vectors and neighbors are, respectively, set to 200 and 10, which are the same as those in [13]. For LDA, NDA, MFA, SDA, LSMDA, and their GRW learning approaches, all (generalized) eigenvectors corresponding to

Table 1
Some examples in the GRW framework.

Classification prototype	Approach				
	LDA	NDA	MFA	SDA	LSMDA
NN	GRW–LDA–NN	GRW–NDA–NN	GRW–MFA–NN	GRW–SDA–NN	GRW–LSMDA–NN
SM	GRW–LDA–SM	GRW–NDA–SM	GRW–MFA–SM	GRW–SDA–SM	GRW–LSMDA–SM
LSM	GRW–LDA–LSM	GRW–NDA–LSM	GRW–MFA–LSM	GRW–SDA–LSM	GRW–LSMDA–LSM

positive (generalized) eigenvalues are preserved to construct the transformation matrices. The free parameters of NDA, MFA, SDA, and LSMDA are selected by 5-fold cross-validation. For the five GRW learning approaches, free parameters can be divided into two groups, of which the first group consists of the parameters of corresponding subspace based feature extraction algorithm, while the second group consists of $\{C,C_e\}$, the parameters of the GRW framework. The overall parameters of these GRW learning approaches are selected in the following hierarchical way: we first select free parameters in the first group by 5-fold cross-validation, next introduce the optimal transformation (that corresponds to the best classification accuracy in this step) into the GRW framework, and then select the free parameters $\{C,C_e\}$ with another independent 5-fold cross-validation. In other words, we treat the GRW framework as a post-processing procedure on the transformation learned by some given feature extraction algorithm, and hence its performance is closely dependent on the results in the first step.

6.1. On synthetic datasets

In the GRW framework (18), there are two free parameters, C and C_e, which, respectively, control the relative importance of the sparse term (the second objective term that leads to sparse weighting coefficients to projection vectors) and that of the shrinking term (the third objective term that shrinks samples within small areas around their homogeneous prototypes). In this section, we will generate synthetic data and study the influence of the above two parameters on classifications. For simplicity, we just test the performance of the GRW–LSMDA–NN approach. The following criterion is employed to evaluate the classification performance:

$$\text{error rate (ER)} = \frac{\#\text{ samples wrongly classified}}{\#\text{ all samples}} \times 100\% \qquad (20)$$

The synthetic data have 12 dimensions, in which two of them are relevant. The probabilities of the class label $y=-1$ and 1 are equal. If $y=-1$, then the first two relevant dimensions $\{x_1, x_2\}$ are drawn from two Gaussian distributions with equal prior probability $p_{11}=p_{12}$, with $\mu_{11}=[6;0]$, $\mu_{12}=[-6;0]$, and $\Sigma_{11}=\Sigma_{12}=[3{,}0;0{,}6]$. If $y=1$, then the first two relevant dimensions $\{x_1,x_2\}$ are drawn from a single Gaussian distribution, with $\mu_2=[0;1]$ and $\Sigma_2=[6{,}0;0{,}0.6]$. For both $y=-1$ and 1, the remaining 10 dimensions are Gaussian noises $x_i \sim N(0{,}64)$, with $i=3,\ldots,12$.

To effectively evaluate the classification performance, we generate 20, 40, 60, 80, 100, 120, 140, 160, 180, and 200 samples per class as the training set, and for each training size, the above generation is repeated 50 times to obtain different training sets. A fixed testing set with 400 samples per class, independently drawn from the same distribution, is adopted to execute error estimation. The experimental results of 50 random generations are averaged. When performing cross-validation, the values of C and C_e are both chosen from the candidate set $\{0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10{,}000\}$. The influence of parameters is evaluated as follows. A set of $\{C, C_e\}$ is determined using cross-validation to obtain the best classification result first, fix one parameter and alter the target parameter, and observe the variation of performances with respect to the target parameter; next change the orders of fixing parameter and target parameter, and then observe the corresponding variation as well. The experimental results are shown in Fig. 2, where x-axis denotes the value of target parameter while y-axis denotes the classification error.

Through Fig. 2, it is noted that for parameter C, most of the best classification results are achieved when C ranges from 0.0001 to 0.01, while for parameter C_e, most of the best results are achieved when C_e equals 0.1. Despite the simplicity of experiments, we did this test only for the variation of one parameter; however, more or less, through Fig. 2 we may obtain some indication that reflects the relative importance of three objective terms. It is shown that the marginal term (the first objective term that leads to large classification margins) plays the most important for classification, the shrinking term the second, and the sparse term the third. The possible reason is explained as follows: compared with the sparse term, the marginal term and the shrinking term contribute more directly to performance improvement; hence they should attract more attention naturally; by comparing the marginal term and the shrinking term, we find that the former aims to obtain a large classification margin, which mainly copes with samples near the classification boundary, while the latter aims to shrink all samples within small areas around their homogeneous prototypes, which copes with all samples, and consequently, in view of enhancing discriminability, adopting the marginal term is more favorable, since moving samples far away from the classification boundary towards their homogeneous prototypes is not only helpless for classification but also prone to overfittings. With an increase of training samples, a continuous performance improvement arises. When the training number per class is larger than 140, this improvement becomes not so obvious; perhaps due to sufficient training samples that have already been applicable, an increase of training samples cannot provide additional discriminative information any more.

In the above paragraphs we have compared the relative importance of three objective terms, and next we concentrate on the sparse objective term and see its effectiveness. In Fig. 3, we plot the weighting coefficients of projection vectors learned by the GRW–LSMDA–NN approach, with respect to different training numbers. The effectiveness of the sparse term is shown intuitively in Fig. 3, and along with the increase of training numbers, this effectiveness becomes more and more obvious. It is noted that when the training number per class is larger than 160, there are always two dominant coefficients, which is consistent with our generative model of synthetic data (the synthetic data are generated by connecting two relevant dimensions and ten noisy dimensions) and further justifies the effectiveness of the sparse term.

6.2. On UCI benchmark datasets

In this section, we conduct experiments on nine UCI benchmark datasets [21]. The overall properties of selected datasets are shown in Table 2.

Among the above datasets, Spectf and Ann-throid have been separated into training set and testing set in advance. For the remaining seven datasets, we need to separate them manually. We perform our separating work as follows: the data are split randomly into a training set consisting of 70% of the whole set and a testing set consisting of the remaining 30%. To control the variability caused by random splittings, the splitting work is repeated 50 times and the resulting accuracies are averaged. Moreover, standard deviations are calculated to describe the uncertainty caused by random splittings. The error rates of all nine datasets are calculated according to Eq. (20).

The average classification error rates and standard deviations are shown in Table 3. For the Pima, Housing, and Spectf datasets, as the number of available LDA projection vectors is only one $(c-1=1)$, the experimental results of GRW–LDA–NN on these three datasets are the same as those of LDA, and we do not give them explicitly to avoid repetitions. LSMDA achieves favorable classification results and outperforms other competing algorithms

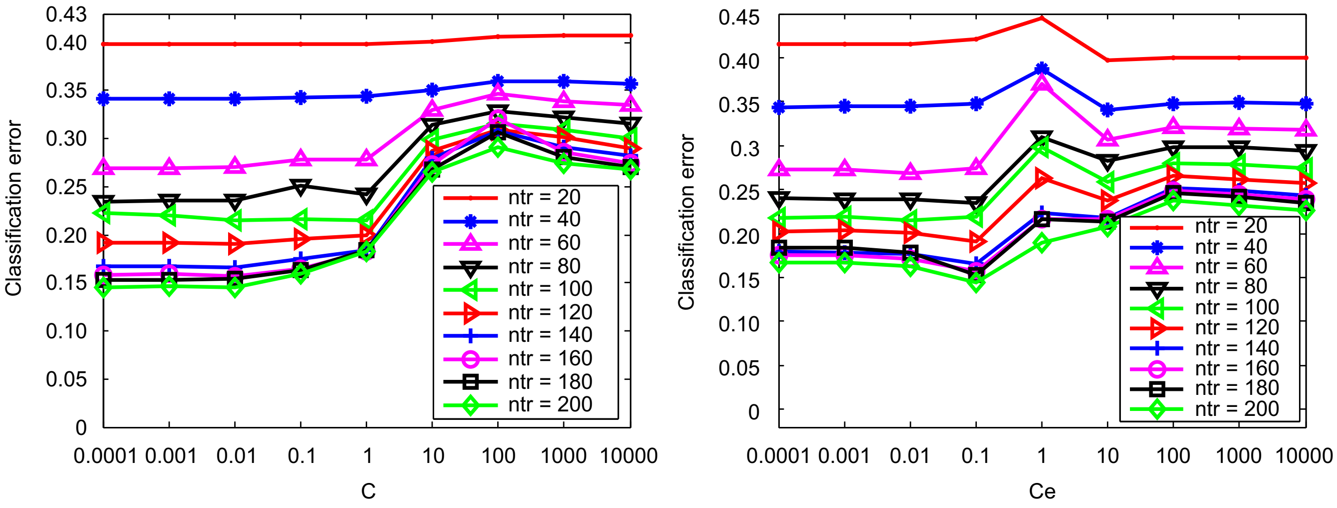

Fig. 2. Influence of parameters C and C_e on classification performance.

on most datasets (except for the Protein dataset). Possible reasons for the wide success of LSMDA might be that compared with other competing algorithms, LSMDA takes most possible application cases into consideration and has fewer application limitations: (1) LSMDA attempts to make up for LDA's first three drawbacks; (2) the within-class scattering matrix of NDA is the same as that of LDA and requires homogeneous samples to be unimodal distributed, whereas LSMDA does not have this requirement; (3) MFA and SDA treats all discrepant vectors equivalently, whereas LSMDA considers their difference; (4) the ability of SDA depends on the clustering performance of the NN-clustering algorithm [12], whereas LSMDA does not have this dependence. In Table 4, we give the average subspace dimensionality of different feature extraction algorithms for 50 random splittings. It is shown that LSMDA gets the highest average subspace dimensionality on all the nine datasets, NDA being the second, MFA the third, SDA the fourth, and LDA the lowest on most datasets. Higher dimensional subspaces may contain more discriminative information, and indeed there is a trend that LSMDA performs the best, MFA and SDA moderate, and LDA the worst, on some of the above datasets. Note that this trend is not suitable to NDA—although NDA has much higher subspace dimensionality than LDA, its classification performance is comparable to that of LDA, and sometimes even worse than that of LDA. This phenomenon might be caused by the mismatch of NDA's within-class scattering matrix with respect to the between-class one, taking on the following two aspects: firstly, NDA's within-class scattering matrix adopts sampling mean as the classification model, but its between-class scattering matrix adopts nearest neighbor (exactly, the average of k nearest neighbors); secondly, NDA's between-class scattering matrix considers the difference of different discrepant vectors, but its within-class scattering matrix does not consider this problem. Additionally, even except for NDA, this trend is not absolute as well, because it is not suitable for datasets that contain too many noisy (at least redundant and useless) components. For the example of the Protein dataset, LDA's average subspace dimensionality is the lowest, but its classification accuracy is the highest. One possible reason might be that there are too many noisy components in this dataset, which leads to a consequence that higher dimensional subspaces not only cannot provide additional discriminative information, but also introduce a large amount of noises that may degrade the classification accuracies. Sometimes we may fall into a dilemma, since a higher dimensional subspace may bring additional discriminative information and noises simultaneously. The GRW framework, which learns the optimal norm of projection vectors, may help us to cope with this dilemma, since through GRW learning, weighting coefficients that correspond to noisy components can be suppressed, and the unfavorable affections caused by noisy components can be weakened as well. It is shown that GRW–LSMDA–NN performs slightly better than GRW–LDA–NN, which displays the effectiveness of the GRW framework for reducing noises and justifies the above point of view. Through a comparison of GRW learning approaches with ODPP, we find that most GRW learning approaches outperform ODPP on most cases, perhaps owing to the following reasons: the GRW framework performs feature weighting on projection vectors, and hence may have more degrees of freedom than ODPP's feature selection on candidate projections. Moreover, through the GRW framework, all related feature extraction algorithms obtain some extent performance improvement, which depicts the general success of this framework for enhancing discriminability.

6.3. *On radar HRRP dataset*

Radar target HRRP is the amplitude of coherent summations of complex time returns from target scatterers in each range resolution cell, which represents the projection of complex returned echoes from the target scattering centers onto the radar line-of-sight (LOS). It contains the target structure signatures, such as target size, scatterer distribution, etc., and therefore, it is a promising signature for radar automatic target recognition (ATR) [22–24]. The target HRRPs are highly variable with target orientation [25], thereby they can be seen as multimodal distributed.

As described in Table 5, the data used to evaluate the recognition performance were measured by a c-band radar with the bandwidth of 400 MHz. The radar HRRP data of three airplanes, including An-26, Yark-42, and Cessna Citation S/II, were measured continuously when the targets were flying. The projections of airplane target trajectories onto the ground are shown in Fig. 4, from which the aspect angle of an airplane can be estimated in virtue of its relative position to radar. Shown in Fig. 5 is an example HRRP waveform. Training set and testing set were from different data segments, among which the 2nd and the 5th segments of Yark-42, the 5th and the 6th segments of An-26, and the 6th and the 7th segments of Cessna Citation S/II, were chosen as the training set (2800 HRRPs), and other data segments were chosen as the testing set (1200 HRRPs). In terms of the

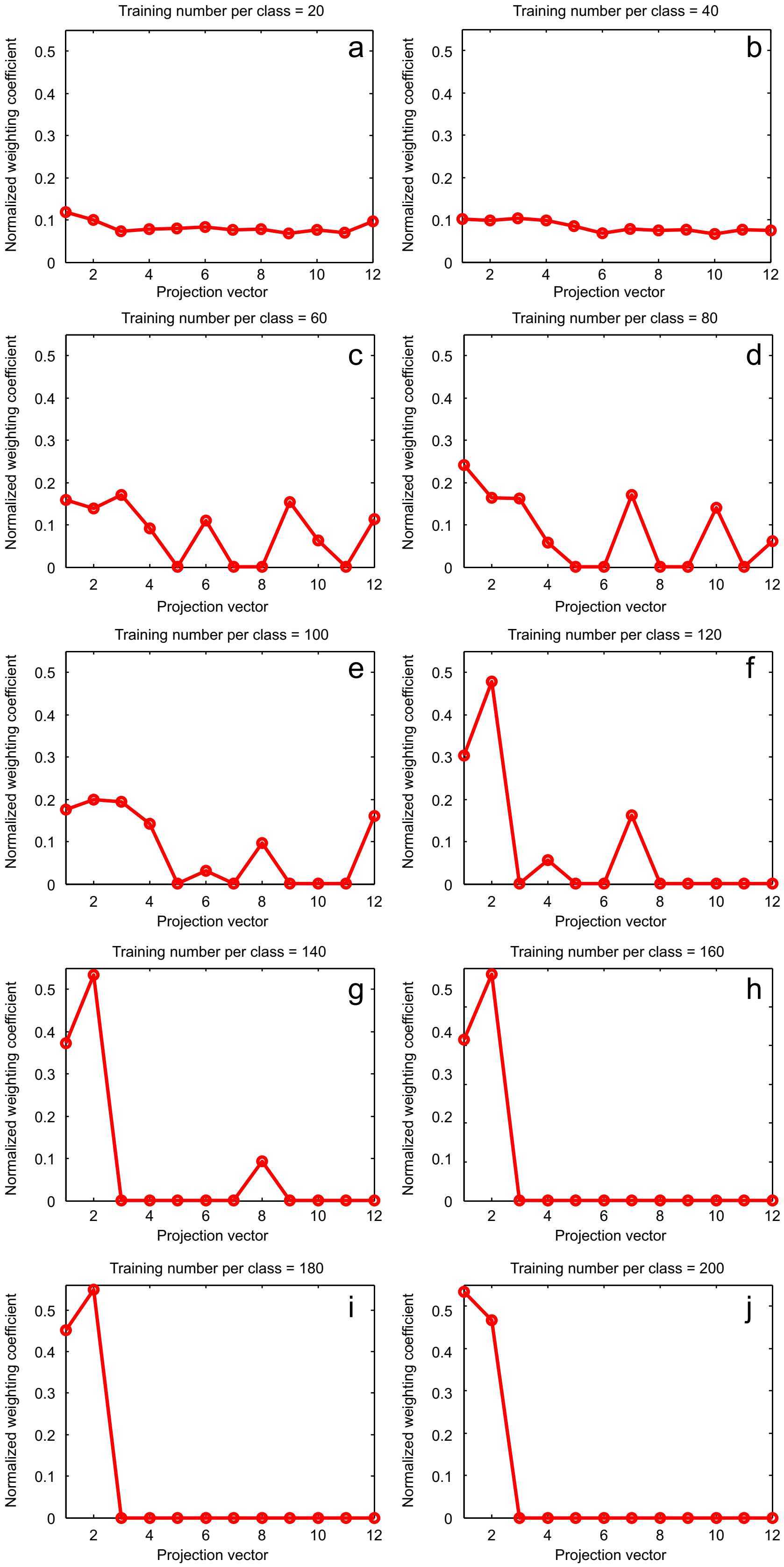

Fig. 3. Weighting coefficients of projection vectors for GRW–LSMDA–NN. From (a) to (j), training number per class ranges from 20 to 200, respectively. The weighting coefficients are normalized through l_1-norm.

requirements for radar HRRP ATR [22–24], the training set should cover almost all of the target-aspect angles of the testing set. According to Fig. 4, this can be satisfied. Moreover, the targets' orientations corresponding to the testing set and training set are different; thus the generalized performances of recognition algorithms can be tested. Additionally, because of the property of time-shift alignment of HRRP [25], the l_2-norm normalized power spectrum of HRRP was used to perform classification because of its time-shift invariance. The power spectrum of HRRP is symmetrical real bilateral spectra, and therefore, it is enough to use half of all features. That is to say each power spectrum for an HRRP sample is a 128-dimensional vector. An example of HRRP power spectrum is shown in Fig. 6. According to radar signal theories, radar HRRP contains many noises and so does its power spectrum [23,26]; thereby it is quite necessary to reduce noisy components and extract discriminative ones. As a result, a classification performance improvement by the LSMDA algorithm and the GRW framework is expected to be obtained.

In the radar HRRP ATR task, we emphasize on the average of the recognition rate for each target. Therefore, we use the balanced error rate (BER) to evaluate the classification performance, and a low BER can ensure that the radar recognition system will not miss any target rather than be partial to some one due to the number of samples:

$$\text{balanced } error \text{ } rate \text{ (BER)}$$
$$= \frac{1}{c}\left(\sum_{i=1}^{c} \frac{\#\text{ samples wrongly classified in the } i\text{th target class}}{\#\text{ samples in the } i\text{th target class}}\right) \times 100\% \tag{21}$$

Table 2
Summary of selected UCI datasets.

Dataset	Size	Dimension	Class
Pima	768	8	2
Housing	506	13	2
Spectf	267	44	2
Ann-throid	7220	21	3
Cmc	1473	9	3
Vehicle	846	18	4
Glass	214	9	6
Dermatology	366	33	6
Protein	116	20	6

Table 3
Classification results on UCI datasets.

Approach	Average error rate (/standard deviation)								
	Pima	Housing	Spectf	Ann-throid	Cmc	Vehicle	Glass	Dermatology	Protein
LDA	0.3019 (/0.0273)	0.1958 (/0.0281)	0.3209	0.0677	0.5576 (/0.0216)	0.2600 (/0.0251)	0.3928 (/0.0657)	0.0406 (/0.0163)	0.2297 (/0.0719)
NDA	0.3043 (/0.0204)	0.1479 (/0.0263)	0.4439	0.0636	0.5587 (/0.0195)	0.2294 (/0.0235)	0.3295 (/0.0548)	0.0409 (/0.0141)	0.2458 (/0.0832)
MFA	0.3145 (/0.0279)	0.1877 (/0.0341)	0.3636	0.0639	0.5567 (/0.0175)	0.2251 (/0.0265)	0.3403 (/0.0544)	0.0559 (/0.0221)	0.3297 (/0.0853)
SDA	0.3090 (/0.0261)	0.1873 (/0.0354)	0.2995	0.0621	0.5603 (/0.0204)	0.2532 (/0.0246)	0.3639 (/0.0724)	0.0355 (/0.0186)	0.2471 (/0.0730)
LSMDA	0.2991 (/0.0236)	0.1473 (/0.0261)	0.1818	0.0315	0.5477 (/0.0223)	0.2191 (/0.0210)	0.2744 (/0.0579)	0.0351 (/0.0155)	0.2452 (/0.0798)
GRW–LDA–NN				0.0657	0.5452 (/0.0203)	0.2458 (/0.0230)	0.3590 (/0.0537)	0.0316 (/0.0125)	0.1955 (/0.0688)
GRW–NDA–NN	0.2823 (/0.01950	0.1226 (/0.0248)	0.2888	0.0508	0.5300 (/0.0177)	0.2028 (/0.0171)	0.2944 (/0.0519)	0.0249 (/0.0108)	0.2032 (/0.0780)
GRW–MFA–NN	0.2790 (/0.0233)	0.1444 (/0.0268)	0.3590	0.0616	0.5303 (/0.0183)	0.2071 (/0.0234)	0.2616 (/0.0404)	0.0428 (/0.0202)	0.2400 (/0.0576)
GRW–SDA–NN	0.3087 (/0.0266)	0.1687 (/0.0337)	0.2941	0.0610	0.5419 (/0.0185)	0.2331 (/0.0223)	0.3256 (/0.0610)	0.0275 (/0.0150)	0.1987 (/0.0706)
GRW–LSMDA–NN	0.2782 (/0.0180)	0.1164 (/0.0236)	0.1551	0.0257	0.5254 (/0.0193)	0.1924 (/0.0176)	0.2502 (/0.0499)	0.0243 (/0.0128)	0.1916 (/0.0706)
ODPP	0.3082 (/0.0292)	0.1879 (/0.0340)	0.2193	0.0665	0.5220 (/0.0215)	0.3203 (/0.0327)	0.2767 (/0.0618)	0.0378 (/0.0170)	0.2852 (/0.0691)

Table 4
Average subspace dimensionality of different feature extraction algorithms.

Approach	Average subspace dimensionality								
	Pima	Housing	Spectf	Ann-throid	Cmc	Vehicle	Glass	Dermatology	Protein
LDA	1	1	1	2	2	3	5	5	5
NDA	7.98	13	29	19	9	18	9	33	20
MFA	6.48	11.76	32	5	8.3	10.1	8.98	17.54	20
SDA	1.12	3.82	5	2	6.64	7.56	5	10.64	7.88
LSMDA	8	13	44	21	9	18	9	33	20

Table 5
Parameters of planes and radar in the radar measured HRRP data experiment.

Radar parameters	Transmitted signal waveform		Chirp signal
	Center frequency		5520 MHz
	Bandwidth		400 MHz
	Pulse repetition frequency		400 Hz
	Sampling frequency after Dechirp		10 MHz
Planes	Length (m)	Width (m)	Height (m)
Yark-42	36.38	34.88	9.83
An-26	23.80	29.20	9.83
Cessna Citation S/II	14.40	15.90	4.57

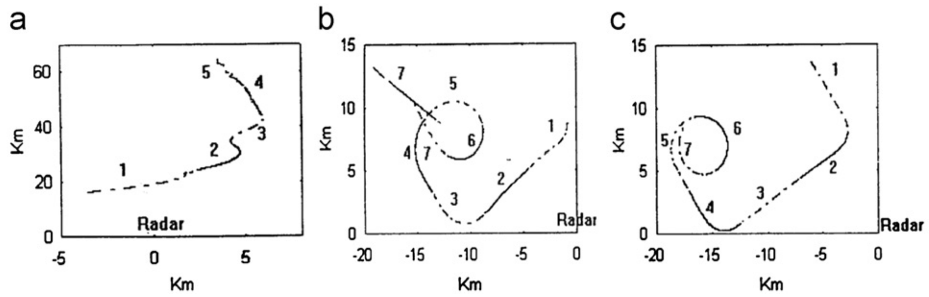

Fig. 4. Projections of plane target trajectories onto the ground: (a) Yark-42, (b) An-26, and (c) Cessna Citation S/II.

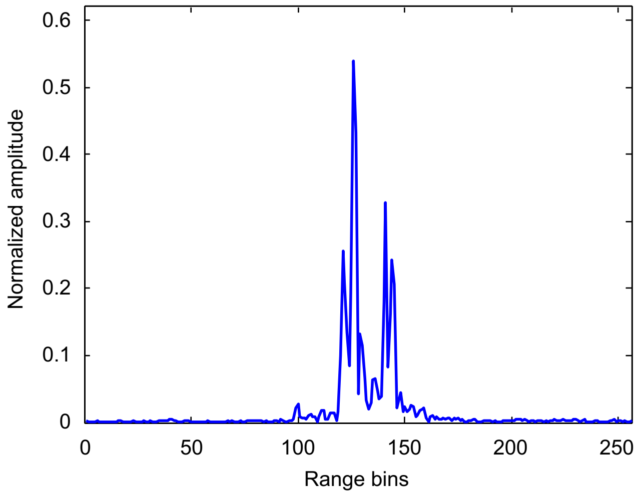

Fig. 5. Waveform of an HRRP example.

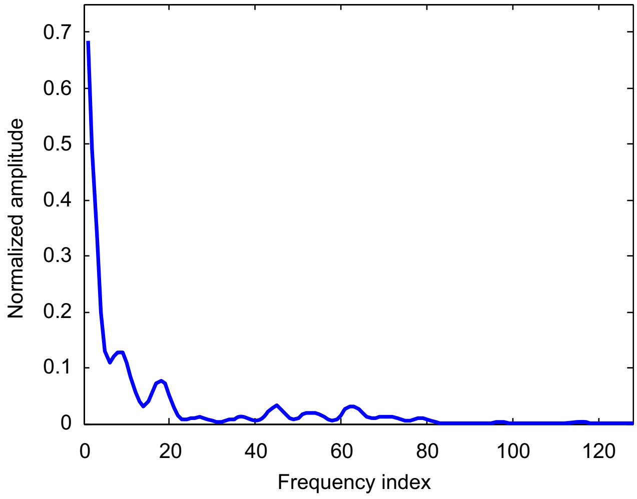

Fig. 6. Power spectrum of an HRRP example.

To effectively compare the performances of different approaches in radar HRRP ATR, we select 20%, 40%, 60%, 80%, and 100% of the training set as training data, and then use all of the testing set as testing data to evaluate performances. For each training percentage, training data are sampled with equal time steps from the whole training set. Experimental results and the corresponding subspace dimensionality are shown in Tables 6 and 7, respectively. Looking through Table 6, it seems that LSMDA cannot obtain satisfactory discriminative ability, since LSMDA performs consistently worse than LDA, MFA, SDA, and ODPP on all training percentages, and just slightly better than NDA on the last three training percentages. However, GRW–LSMDA–NN outperforms all other competing approaches on all training percentages, which seems to conflict with above classification results. This seeming confliction can be explained as follows: radar HRRP power spectrum contains many noisy components (according to radar signal theories [23,26] and wavelet compressive theories [27,28]; most target information is included in the low-frequency band, while in the high-frequency band there exists a large amount of noises), and LSMDA consistently results in high dimensional transformation subspaces (the subspace dimensionality of LSMDA equals 128 on all training percentages, which is typically larger than that of other competing feature extraction algorithms, see Table 7); as a result, some projection vectors of LSMDA correspond to noisy components, and the existence of these projection vectors may degrade the classification performances of LSMDA; through the GRW learning, we can weaken the disadvantageous affections derived from unfavorable projection vectors by endowing small weighting coefficients on these projection vectors, and hence, may help to improve LSMDA's discriminative abilities. In short, the combination of LSMDA and the subsequent GRW learning framework provides us with a promising tool to cope with the radar HRRP ATR problem.

Table 6
Classification results on radar HRRP dataset.

Approach	Error rate via training percentage				
	20%	40%	60%	80%	100%
LDA	0.1542	0.1333	0.0892	0.0758	0.0625
NDA	0.1869	0.1790	0.1576	0.1545	0.1492
MFA	0.1542	0.1517	0.1281	0.1192	0.0958
SDA	0.1600	0.1517	0.1250	0.1142	0.0867
LSMDA	0.1900	0.1858	0.1549	0.1508	0.1308
GRW–LDA–NN	0.1500	0.1192	0.0883	0.0717	0.0592
GRW–NDA–NN	0.1417	0.1358	0.0950	0.0892	0.0725
GRW–MFA–NN	0.1358	0.1208	0.1017	0.0783	0.0683
GRW–SDA–NN	0.1283	0.1250	0.0983	0.0925	0.0767
GRW–LSMDA–NN	0.1233	0.1133	0.0758	0.0625	0.0492
ODPP	0.1479	0.1342	0.0886	0.0709	0.0603

Table 7
Subspace dimensionality of different feature extraction algorithms.

Approach	Subspace dimensionality via training percentage				
	20%	40%	60%	80%	100%
LDA	2	2	2	2	2
NDA	41	87	103	128	128
MFA	20	28	46	57	81
SDA	8	14	14	14	14
LSMDA	128	128	128	128	128

7. Conclusions and future work

According to the analysis of intrinsic drawbacks of traditional LDA, in this paper, we propose a LSMDA feature extraction algorithm and a GRW framework, respectively, to make up for the first three and the fourth drawbacks of LDA. In LSMDA, a new model, local sampling mean, is employed as the basic classification prototype to construct scattering matrices. Through weighting discrepant vectors according to corresponding sample's classifiability, our new scattering matrices take it into full consideration that different discrepant vectors have different contributions to classifications. The GRW framework, which performs as a post-processing procedure on the learned transformation, is proposed to learn the optimal norm of projection vectors. Furthermore, the GRW framework can be utilized for any subspace based feature extraction algorithm and any prototype based classifier, which makes it very general in real applications.

Although the LSMDA algorithm and GRW framework perform well in most experiments, there are still some open problems that may restrict the applications of our work and need to be solved in the future. First, when applying to datasets that contain heavily corrupted samples or outliers, LSMDA and the GRW framework may perform poorly. Thereby, how to detect these samples and integrate the detection into our work is an urgent problem for real applications. Second, there are overall four free parameters (σ_1, σ_2, C, and C_e) in our work, and selecting them through cross-validation is a time-consuming task. However, if we can obtain some useful prior information (e.g., data structure, data distribution, proportion of noises, etc.), we may have an expectation to simplify the parameter selection task. Consequently, how to design simpler techniques to more easily select free parameters is still a challenging task in the future.

Acknowledgements

The authors are grateful to the National Laboratory of Radar Signal Processing for providing the radar HRRP data. This work is partially supported by the Program for Cheung Kong Scholars and Innovative Research Team in University (PCSIRT, Grant no. IRT0645) and NSFC through Grant 60772140.

References

[1] H. Liu, H. Motoda, Feature Extraction, Construction and Selection: A Data Mining Perspective, Kluwer Academic Publishers, Boston, USA, 1998.

[2] I. Guyon, S. Gunn, M. Nikravesh, L.A. Zadeh, Feature Extraction: Foundations and Applications, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[3] I.J. Jolliffe, Principal Component analysis, Springer-Verlag, New York, 1986.

[4] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 711–720.

[5] B. Scholkopf, A. Smola, K.R. Muller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Computation 10 (5) (1998) 1299–1319.

[6] G. Baudat, F. Anouar, Generalized discriminant analysis using a kernel approach, Neural Computation 12 (10) (2000) 2385–2404.

[7] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd edition, Academic Press, Boston, USA, 1990.

[8] M. Bressan, J. Vitria, Nonparametric discriminant analysis and nearest neighbor classification, Pattern Recognition Letters 24 (15) (2003) 2743–2749.

[9] Z.F. Li, W. Liu, D.H. Lin, X.O. Tang, Nonparametric subspace analysis for face recognition, in: Proceedings of the International Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 2005, pp. 961–966.

[10] S.C. Yan, D. Xu, B.Y. Zhang, H.J. Zhang, Q. Yang, S. Lin, Graph embedding and extensions: a general framework for dimensionality reduction, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (1) (2007) 40–51.

[11] H.T. Chen, H.W. Chang, T.L. Liu, Local discriminant embedding and its variants, in: Proceedings of the International Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 2005, pp. 846–853.

[12] M. Zhu, A.M. Martinez, Subclass discriminant analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (8) (2006) 1274–1286.

[13] Y. Su, S.G. Shan, X.L. Chen, W. Gao, Classifiability-based optimal discriminatory projection pursuit, in: Proceedings of the International Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska, USA, 2008, pp. 23–28.

[14] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences 55 (1) (1997) 119–139.

[15] Y. Ma, Y. Ijiri, S. Lao, M. Kawade, Re-weighting linear discrimination analysis under ranking loss, in: Proceedings of the International Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska, USA, 2008, pp. 1–8.

[16] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, John Wiley and Sons, Inc., New York, 2001.

[17] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, Cambridge, UK, 2004.

[18] K.Q. Weinberger, J. Blitzer, L.K. Saul, Distance metric learning for large margin nearest neighbor classification, in: Neural Information Processing Systems, 2006, pp. 1473–1480.

[19] C.J. Veenman, D.M.J. Tax, LESS: a model-based classifier for sparse subspaces, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (9) (2005) 1496–1500.

[20] B. Chen, H. Liu, J. Chai, Z. Bao, Large margin feature weighting method via linear programming, IEEE Transactions on Knowledge and Data Engineering 21 (10) (2009) 1475–1488.

[21] C.L. Blake, C.J. Merz, UCI Repository of Machine Learning Databases, Department of Information and Computer Sciences, University of California, Irvine. Available at 〈http://archive.ics.uci.edu/ml/〉.

[22] B. Pei, Z. Bao, Bispectrum based approach to high radar range profile for automatic target recognition, Pattern Recognition 35 (11) (2002) 2643–2651.

[23] L. Du, H. Liu, Z. Bao, M. Xing, Radar HRRP target recognition based on higher order spectra, IEEE Transactions on Signal Processing 53 (7) (2005) 2359–2368.

[24] L. Du, H. Liu, Z. Bao, J. Zhang, A two-distribution compounded statistical model for radar HRRP target recognition, IEEE Transactions on Signal Processing 54 (6) (2006) 2226–2238.

[25] M. Xing, Z. Bao, B. Pei, Properties of high-resolution range profiles, Opt. Eng 41 (2) (2002) 493–504.

[26] Z. Bao, M. Xing, T. Wang, Radar Imaging Technique, Publishing House of Electronics Industry, 2005.

[27] S. Mallat, A Wavelet Tour of Signal Processing, Academic Press, 1998.

[28] I. Daubechies, Ten Lectures on Wavelets, SIAM, 1992.

JING CHAI received his B.Eng. degree in electronic engineering from Xidian University in 2005. He is currently working towards the Ph.D. degree with the National Laboratory of Radar Signal Processing, Xidian University. His research interests include radar signal processing, radar automatic target recognition, and machine learning.

HONGWEI LIU received his Ph.D. degree in Xidian University in 1999. From 2001 to 2002, he was a visiting scholar at Duke University, USA. He currently holds a professor position and is the director of Key Laboratory of Radar Signal Processing, Xidian University. His research interests are radar automatic target recognition and radar signal processing.

ZHENG BAO is a professor at Xidian University and an academician of the Chinese Academy of Science. He is the author or co-author of six books and has published more than 300 papers. Now his research work focuses on the areas of space–time adaptive processing, radar imaging, and radar automatic target recognition.