IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 1, 2018, accepted November 10, 2018, date of publication November 14, 2018, date of current version December 19, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2881256

Reversible Discriminant Analysis

LAN BAI[1], ZHEN WANG[1], YUAN-HAI SHAO[2], AND CHUN-NA LI[3]
[1]School of Mathematical Sciences, Inner Mongolia University, Hohhot 010021, China
[2]School of Economics and Management, Hainan University, Haikou 570228, China
[3]Zhijiang College, Zhejiang University of Technology, Hangzhou 310024, China

Corresponding author: Yuan-Hai Shao (shaoyuanhai21@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 11501310, Grant 61866010, Grant 11871183, and Grant 61703370, in part by the Inner Mongolia Autonomous Region University Scientific Research Project under Grant NJZC17006, in part by the Natural Science Foundation of Hainan Province under Grant 118QN181, in part by the Scientific Research Foundation of Hainan University under Grant kyqd(sk)1804, and in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LY15F030013 and Grant LQ17F030003.

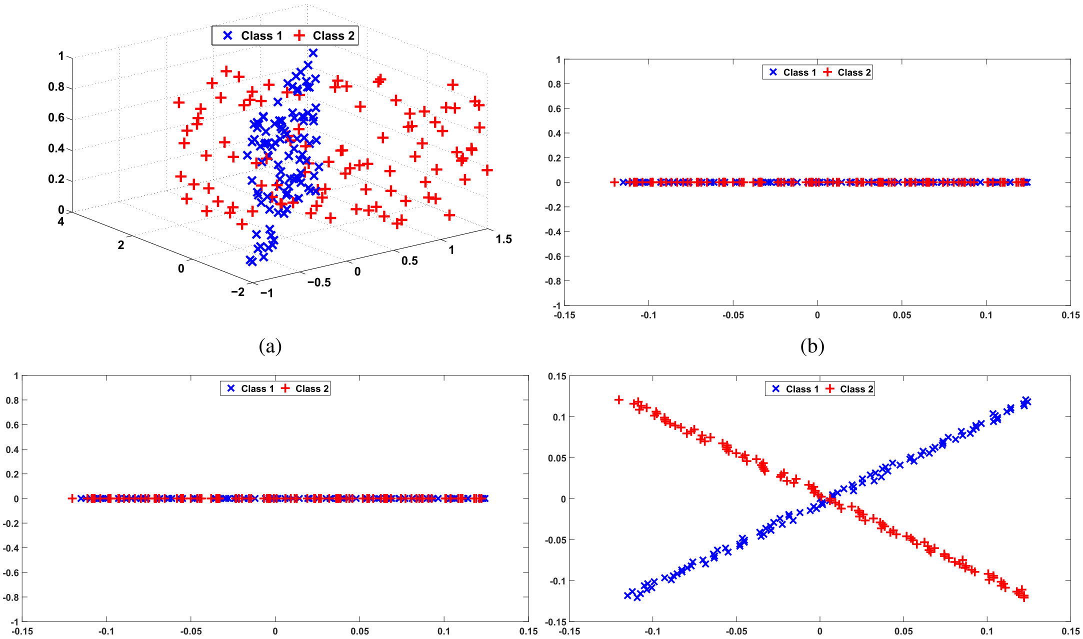
ABSTRACT Principal component analysis (PCA) and linear discriminant analysis (LDA) have been extended to be a group of classical methods in dimensionality reduction for unsupervised and supervised learning, respectively. However, compared with the PCA, the LDA loses several advantages because of the singularity of its between-class scatter, resulting in singular mapping and restriction of reduced dimension. In this paper, we propose a dimensionality reduction method by defining a full-rank between-class scatter, called reversible discriminant analysis (RDA). Based on the new defined between-class scatter matrix, our RDA obtains a nonsingular mapping. Thus, RDA can reduce the sample space to arbitrary dimension and the mapped sample can be recovered. RDA is also extended to kernel based dimensionality reduction. In addition, PCA and LDA are the special cases of our RDA. Experiments on the benchmark and real problems confirm the effectiveness of the proposed method.

INDEX TERMS Between-class scatter, dimensionality reduction, linear discriminant analysis, supervised learning.

I. INTRODUCTION

Principal component analysis (PCA) [1], [2], the classical linear dimensionality reduction method for unsupervised learning, has been widely studied and applied [3]–[7]. PCA seeks an orthogonal mapping such that the mapped samples are as far as possible to each other, leading to solve an eigenvalue problem. The dimension of the sample space can be reduced by the eigenvectors corresponding to some larger eigenvalues, and the principle component of the mapping can be estimated by the sum of the eigenvalues corresponding to the selected eigenvectors. It is easy to obtain a nonsingular mapping by setting the principle component to 100% in PCA, i.e., the original samples can be recovered without loss of information. Generally, PCA does not suit for classification, because PCA ignores the information from the classes. In contrast, linear discriminant analysis (LDA) [8], [9], another classical dimensionality reduction method, is proposed for supervised learning. LDA hires the within-class scatter and between-class scatter such that the mapped samples from the same class are close to the class center and the class centers are far away from the other class centers, leading to solve a generalized eigenvalue problem (GEP). LDA has also been widely studied and applied [10]–[18].

Instead of a nonsingular mapping by PCA, LDA would obtain a singular mapping, because its between-class scatter matrix may be singular. Some improvements were proposed to obtain nonsingular mapping, e.g., orthogonal least squares discriminant analysis (OLSDA) [11], orthogonal centroid method (OCM) [13], minimal distance maximization (MDM) [15], maximum margin criterion (MMC) [14], fisher discriminant analysis with L1-norm (L1LDA) [10], worst-case linear discriminant analysis (WLDA) [16], and linear discriminant analysis with worst between-class separation and average within-class compactness (WSAC) [17]. However, OLSDA ignores the between-class scatter, OCM and MDM ignores the within-class scatter. They don't use the class information sufficiently. MMC replaces division in LDA with substraction to avoid the computation of the inverse of a matrix, but it may obtain some mapping vectors corresponding to negative eigenvalues, resulting in trouble for classification. L1LDA that reduces one dimension by solving a GEP in each iteration, WLDA and WSAC that solve many semi-definite



(a) (b)

(c) (d)

FIGURE 1. Toy example: (i) original dataset includes 202 samples in R^3, where '×' denotes class 1 and '+' denotes class 2; (ii) dimension redundancy samples obtained by LDA in R; (iii) dimension redundancy samples obtained by RDA in R; (iv) dimension redundancy samples obtained by RDA in R^2. (a) Dataset. (b) LDA. (c) RDA in R. (d) RDA in R^2.

programming problems, are often sensitive for the initialization and spend too much learning time.

In this paper, we propose a reversible discriminant analysis (RDA) for dimensionality reduction. Similar to the LDA, our RDA considers the within-class compactness and between-class separation. However, a full-rank between-class scatter matrix is defined and used in RDA, which makes the RDA obtain a nonsingular mapping. Thus, our RDA can reduce the sample space to arbitrary dimension, and the primal space can be recovered by the inverse of the mapping. In fact, the between-class scatter used in our RDA includes three different scatters, i.e., sample-to-sample, class-to-class, and sample-to-class. The sample-to-sample scatter used in the PCA keeps the mapped samples discriminative to each other. The class-to-class scatter used in the LDA keeps the mapped class centers discriminative to each other. And the sample-to-class scatter, which has not been used for dimensionality reduction, keeps the mapped samples discriminative to different class centers. Therefore, the PCA and the LDA are the special cases of the RDA. Our RDA solves a GEP similar to the LDA, and it obtains a generalized orthogonal mapping, which can be regarded as an orthogonal mapping in a new space. Thus, the mapping vectors in the RDA can be selected by their corresponding eigenvalues similar to PCA.

Now, we give an example to show the superiority of the full rank between-class scatter used in RDA. In Fig. 1(i), there are about two hundreds samples in R^3 from binary classes. Fig. 1(ii) shows the samples after reducing dimension by LDA. Due to the mapping space's dimension is no more than $k-1$ by LDA, where k is the class number, it is obvious that LDA reduces the dimension of these samples to one dimension, resulting in difficulty for classification. Figs. 1 (iii) and (iv) show the results of our RDA on the toy example. It is obvious that the samples by RDA overlap in R but distinguish to each other in R^2. Thus, the samples in Fig. 1(iv) can be classified much better than in Fig. 1 (ii) or (iii) by some classification methods, e.g., GEPSVM [19] or TWSVM [20], [21]. In addition, we also extend RDA to kernel based dimensionality reduction. The experimental results on benchmark and practical problems show its better performance compared with LDA and its extensions.

The rest of this paper is organized as follows: Section II briefly reviews PCA, LDA and its extensions; in Section III, we elaborate the RDA; the experiments are arranged in Section IV; finally, some conclusions are given.

II. BACKGROUND

Consider a dataset in the n-dimensional real space R^n represented by $X = (x_1, x_2, \ldots, x_m) \in R^{n\times m}$, where $x_j \in R^n$ is the jth sample with its label $y_j \in \{1, 2, \ldots, k\}$. Let c_i ($i = 1, \ldots, k$) be the center of the ith class. Without loss of generality, suppose the samples are centeralized, i.e., the whole data center is at origin. Below, we give a brief outlines of PCA, LDA, OLSDA, OCM, and MMC.

A. PCA

PCA [1] directly uses X without Y for dimensionality reduction. By the covariance matrix $S = \sum_{i=1}^{m} x_i x_i^\top$, PCA maximizes

$$J_{PCA}(w) = w^\top S w, \tag{1}$$

where $w \in R^n$ is the mapping vector, leading to solve an eigenvalue problem as follow

$$Sw = \lambda w, \tag{2}$$

where $\lambda \in R$ is the eigenvalue. Note that S only considers the relation of the samples without class information, which is regarded as sample-to-sample scatter.

In practice, S is positive definite, and thus there are n linear independent eigenvectors which construct an orthogonal mapping. Therefore, the samples can be reduced to arbitrary dimension and the mapped samples can be recovered. The contribution of each mapping vector (i.e., the eigenvector) can be estimated by its corresponding eigenvalue.

B. LDA

For supervised learning, LDA [8] defines the within-class scatter matrix S_w and the between-class scatter matrix S_b as

$$S_w = \sum_{i=1}^{k} \sum_{j \in N_i} (x_j - c_i)(x_j - c_i)^\top,$$
$$S_b = \sum_{i=1}^{k} m_i c_i c_i^\top, \tag{3}$$

where N_i is the index set of the ith class with $i = 1, \ldots, k$. To realize the within-class compactness and the between-class separation, LDA maximizes the so-called Fisher criterion [22]

$$J_{LDA}(w) = \frac{w^\top S_b w}{w^\top S_w w}, \tag{4}$$

leading to solve a generalized eigenvalue problem as

$$S_b w = \lambda S_w w, \tag{5}$$

where λ is the generalized eigenvalue of S_w w.r.t. S_b, and $w \neq 0$ is its corresponding eigenvector. Note that S_b only considers the relation of the class centers, which is regarded as class-to-class scatter.

In practice, LDA obtains $k-1$ eigenvectors ($k << n$) corresponding to $k-1$ nonzero eigenvalues, because $rank(S_b) = k-1$. Thus, LDA obtains a singular mapping, which reduces the samples into a space with the dimension no more than $k-1$, and the mapped samples cannot be recovered.

C. OLSDA

To obtain a nonsingular mapping in supervised learning, OLSDA [11] considers the within-class compactness only by minimizing

$$J_{OLSDA}(w) = w^\top S_w w, \tag{6}$$

leading to solve an eigenvalue problem as

$$S_w w = \lambda w. \tag{7}$$

The eigenvectors corresponding to smaller eigenvalues are selected as the mapping vectors.

Since S_w is positive definite in practice, OLSDA can obtain an orthogonal mapping to reduce the samples to arbitrary dimension. However, OLSDA ignores the between-class separation which may cause trouble for classification. For example, consider four samples from two classes, where the samples are at the vertexes of a box and the diagonal two samples belongs to the same class. Due to the two classes share the same class center, OLSDA reduces these samples at one point.

D. MMC

Different from LDA uses division to measure the within-class and between-class scatters, MMC [23] uses subtraction to measure them by maximizing

$$J_{MMC}(w) = w^\top S_b w - w^\top S_w w, \tag{8}$$

leading to solve an eigenvalue problem as

$$(S_b - S_w)w = \lambda w. \tag{9}$$

Thus, MMC can obtain an orthogonal mapping.

It is worth to mention that MMC may obtain a much small $w^\top S_b w$ in (8) corresponding to negative eigenvalues from (9). Thus, the mapped class centers would be close to each other by the corresponding eigenvectors, resulting in trouble for classification. By the way, Song *et al.* [24] improved MMC by adding a regular parameter between two parts in (8).

III. RDA

A. LINEAR FORMATION

Before we elaborate our RDA, let us define a new between-class scatter matrix as

$$S_b^* = \sum_{i=1}^{k} \frac{m_i}{m - m_i} \sum_{j \notin N_i} (c_i c_i^\top - \gamma_1 (c_i x_j^\top + x_j c_i^\top) + \gamma_2 x_j x_j^\top), \tag{10}$$

where γ_1 and γ_2 are nonnegative parameters.

It is obvious that the new between-class scatter includes the class-to-class scatter used in S_b in LDA and the sample-to-sample scatter used in S in PCA, which helps RDA to achieve the purposes of LDA and PCA to some extent. The new part, i.e., the symmetric part $(c_i x_j^\top + x_j c_i^\top)$, which we called sample-to-class scatter, is first introduced into the between-class scatter. The geometric interpretation of the between-class scatters is shown in Fig. 2. Maximizing the class-to-class between-class scatter ignores the data structure, while minimizing the sample-to-class one leads the samples from different class lie on the opposite direction of the corresponding class center. Thus, our RDA maximizes

$$J_{RDA}(w) = \frac{w^\top S_b^* w}{w^\top S_w w}. \tag{11}$$

The mapping vectors can be obtained by solving following generalized eigenvalue problem

$$S_b^* w = \lambda S_w w. \tag{12}$$

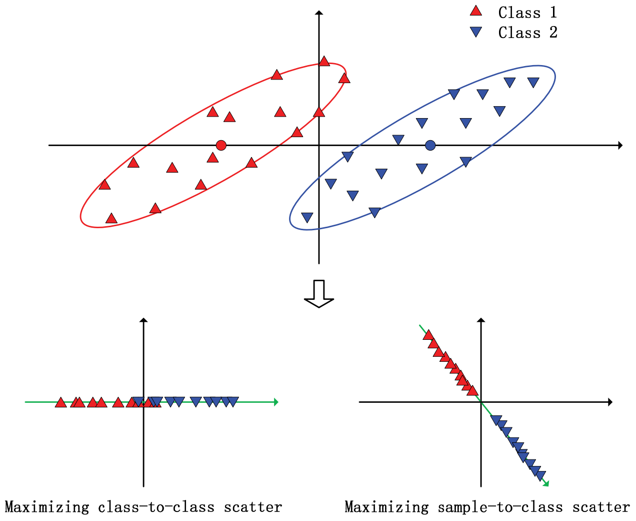

Class 1
Class 2
Maximizing class-to-class scatter
Maximizing sample-to-class scatter

FIGURE 2. Geometric interpretation of maximizing class-to-class and sample-to-class scatter. There are two classes in R^2, where two circles (c_1 and c_2) lie on the horizontal axis denote the class centers, respectively. For between-class scatter, two strategies to reduce the above samples into R are to maximize class-to class scatter (used in LDA) or sample-to class scatter (used in RDA). Maximizing class-to-class scatter ($w^\top||c_1 - c_2||^2 w$) requires the class centers have the largest distance in the reduced feature space, resulting in two classes lies on the horizontal axis and thus they may overlap each other (showed in the bottom-left figure). Maximizing sample-to-class scatter ($-w^\top(x_i^\top c_1)w$ where x_i ∈Class 2, or $-w^\top(x_j^\top c_2)w$ where x_j ∈Class 1) requires the sample and the different class center lie on the opposite directions, resulting in two classes lie on such two opposite directions that they can be classified easily (showed in the bottom-right figure).

Generally, S_w is full-rank. In fact, we have

Theorem 1: S_w is positive definite if and only if there exist n pairs of i and j $(1 \leq i \leq k, j \in N_i)$ such that the vectors $x_j - c_i$ are linear independent.

Proof: Suppose there are n pairs of i and j $(1 \leq i \leq k, j \in N_i)$ such that the vectors $x_j - c_i$ are linear independent. Then, we can construct a matrix $A \in R^{n\times n}$ where its columns are these n linear independent vectors. Since A is nonsingular, and since $S_w = A * A^\top + R$ (where R is the rest component and positive semi-definite), S_w is positive definite. On the contrary, suppose S_w is positive definite. Thus, S_w must have the decomposition $S_w = A * A^\top + R$, where $A * A^\top$ is positive definite. Note that S_w is the sum of $x_j - c_i$ for all $1 \leq i \leq k, j \in N_i$. The columns of A are the n linear independent $x_j - c_i$. □

Instead of $rank(S_b) = k - 1$ in LDA, we have a full-rank S_b^* in RDA in practice. In fact, we have

Theorem 2: Suppose $\gamma_2 \neq 0$, $\gamma_1^2 \leq \gamma_2$, and there exist n pairs of i and j $(1 \leq i \leq k, j \notin N_i)$ such that the vectors $\gamma_1 x_j - c_i$ are linear independent, then S_b^* is positive definite.

Proof: Note that

$$S_b^* = \sum_{i=1}^{k} \frac{m_i}{m-m_i} \sum_{j\notin N_i}((c_i-\gamma_1 x_j)(c_i - \gamma_1 x_j)^\top + (\gamma_2-\gamma_1^2)x_j x_j^\top). \tag{13}$$

From the proof of Theorem 1, it is easy to obtain the conclusion. □

In practical problems that the sample number is much larger than the dimension, the conditions of Theorems 1 and 2 are always satisfied. Thus, RDA obtains a nonsingular linear mapping, which reduces the sample space into a new space with arbitrary dimension. If the sample number is smaller than the dimension, e.g., small sample size problems [25]–[28], the conditions of Theorems 1 and 2 cannot be satisfied, i.e., S_w or S_b^* may be positive semi-definite. However, the regularization technique [29] can help them to be positive definite by adding an ϵI on them, where $\epsilon > 0$ is a very small number and I is the identity matrix. In the following, we discuss the generalized orthogonality of RDA and how to estimate the reconstruction error for a generalized orthogonal mapping.

Note that RDA is equivalent to

$$\begin{aligned} &\max_W tr(W^\top S_b^* W) \\ &s.t.\ W^\top S_w W = I, \end{aligned} \tag{14}$$

where $W \in R^{n\times n}$. The solution of the above problem is such a matrix where its columns are the n generalized eigenvectors to (12). Thus, the mapping obtained by RDA is generalized orthogonal w.r.t. S_w.

Note that S_w has the unique Cholesky decomposition as $S_w = L^\top L$, where L is an upper triangle matrix and its diagonal elements are larger than 0. Let $V = L^{-1}$ and $W = VU$, then the problem (14) is recast to

$$\begin{aligned} &\max_U tr(U^\top V^\top S_b^* VU) \\ &s.t.\ U^\top U = I. \end{aligned} \tag{15}$$

Since the linear mapping V is nonsingular and it is constant for the given samples, each column of W is decided by V and the corresponding column of U, i.e., discarding some columns from W is equivalent to discarding corresponding columns from U. Due to the columns of U correspond to the generalized eigenvalues to (12) (i.e., the eigenvalues to $U^\top V^\top S_b^* VUw = \lambda w$), the mapping vectors should be selected by their corresponding eigenvalues the same as PCA.

Moreover, the process of RDA can be regarded as two steps. First, the primal samples are transformed by a nonsingular linear mapping V. Then, an orthogonal mapping U is used to maximize the between-class scatter in the transformed space. Since the loss of information when discarding some mapping vectors just happens in the second step, the reconstruction error can be estimated by

$$\text{RE} = ||X^\top V - X^\top V \tilde{U} \tilde{U}^\top||_F, \tag{16}$$

where $\tilde{U}$ is constructed by the columns selected from U (corresponding to the columns of W), and $\|\cdot\|_F$ is the Frobenius norm.

In addition, PCA and LDA are the special cases of our RDA by adjusting appropriate parameters. It is obvious that

Theorem 3: (1) If $\gamma_1 = \gamma_2 = 0$, RDA is equivalent to LDA.
(2) If $\gamma_1 = 0$, $\gamma_2 \rightarrow \infty$, RDA is approximate to PCA.

B. KERNEL BASED FORMATION

In this subsection, we extend RDA to kernel based dimensionality reduction [30], [31]. Suppose the data samples $X \in R^{n\times m}$ are first transformed into a high dimensional space $\mathcal{F}$ by an nonlinear mapping $\phi(\cdot)$. Then, RDA is implemented in the space $\mathcal{F}$, i.e., we seek the mapping vector w in $\mathcal{F}$. According to the theory of reproducing kernels [30], the mapping vector $w \in \mathcal{F}$ must lie in the span of all transformed data samples $\phi(X)$, i.e., an expansion of w can be $\sum_{i=1}^{m} \alpha_i\phi(x_i)$, where $\alpha_i \in R$ with $i = 1, \ldots, m$. Thus, we have

$$\begin{aligned} < \phi(x), w > &= \sum_{i=1}^{m} \alpha_i < \phi(x_i), \phi(x) > \\ &= \sum_{i=1}^{m} \alpha_i K(x_i, x) = a^\top K(x, X), \end{aligned} \tag{17}$$

where $< \cdot, \cdot >$ denotes the inner product, $K(\cdot, \cdot)$ is a predetermined kernel function, and all of a_i $(i = 1, \ldots, m)$ construct $a \in R^m$.

Without loss of generality, suppose the samples in X have been collected together by the class labels, i.e., $X = (X_1, \ldots, X_k)$. Thus, the within-class scatter and between-class scatter matrices in $\mathcal{F}$ can be defined as

$$\begin{aligned} S_w^\phi &== \phi(X)D\phi(X)^\top \\ S_b^{*\phi} &= \phi(X)(E - \gamma_1 E' + \gamma_2 M)\phi(X)^\top, \end{aligned} \tag{18}$$

where $D, E, E', M \in R^{m\times m}$. D and E are two block diagonal matrices, where the ith block diagonal element of D is $I - \frac{1}{m_i} e_{m_i} e_{m_i}^\top$, the ith one of E is $(\frac{1}{m_i} + \frac{2\gamma_1}{m-m_i}) e_{m_i} e_{m_i}^\top$, with $i = 1, 2, \ldots, k$, and e_{m_i} is a vector of ones with the dimension m_i. $E'(i,j) = \frac{1}{m-m_{y_i}} + \frac{1}{m-m_{y_j}}$ with $i, j = 1, 2, \ldots, m$. M

TABLE 1. LOO results by NN/SVM classifier with linear dimensionality reduction on the benchmark datasets.

	Baseline	PCA	PCA+LDA	LDA	OLSDA	OCM	MMC	RDA
Data	Acc./Acc.	Acc./Acc.	Acc./Acc.	Acc./Acc.	Acc./Acc.	Acc./Acc.	Acc./Acc.	Acc./Acc.
		Time(Sec.)	Time(Sec.)	Time(Sec.)	Time(Sec.)	Time(Sec.)	Time(Sec.)	Time(Sec.)
m×n (k)		Dim.	Dim.	Dim.	Dim.	Dim.	Dim.	Dim.
Heartc	59.41/87.03	60.07/76.01	59.74/74.79	96.70/97.03	92.74/96.37	65.02/86.14	60.40/97.03	**99.67/99.01**
		0.0002	0.0050	0.0003	0.0001	0.0001	0.0002	0.0001
303×14 (2)		2	1	1	7	2	7	2
Heartstatlog	57.41/81.11	58.89/81.48	55.56/86.30	75.19/84.44	75.56/90.37	65.56/87.78	61.85/84.44	**81.85/91.11**
		0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
270×13 (2)		3	1	1	10	2	2	11
Hepatitis	69.68/86.13	69.68/87.10	67.74/89.35	83.23/79.35	80.00/80.00	69.68/**92.26**	69.68/88.39	**86.45**/89.35
		0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001
155×19 (2)		3	1	1	1	5	3	1
Hourse	69.00/61.33	70.33/64.67	61.00/60.33	71.00/62.33	75.33/76.33	69.33/60.00	70.33/63.67	**82.00/80.33**
		0.0001	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001
300×26 (2)		2	1	1	5	4	2	21
Housevotes	92.41/97.47	92.41/97.47	92.18/93.79	94.48/96.78	95.17/96.78	93.79/97.70	93.33/97.24	**96.32/98.85**
		0.0001	0.0001	0.0004	0.0002	0.0002	0.0003	0.0003
435×16 (2)		16	1	1	6	14	9	8
Ionosphere	86.61/87.32	90.03/90.20	84.05/92.08	82.34/91.51	89.46/90.49	90.31/90.77	90.03/89.19	**95.16/95.03**
		0.0001	0.0001	0.0004	0.0001	0.0001	0.0001	0.0002
351×33 (2)		8	1	1	10	9	11	4
Sonar	82.69/87.98	83.65/87.98	72.12/85.02	71.63/86.13	84.13/87.94	**86.54**/87.02	83.17/87.02	**86.54/88.46**
		0.0001	0.0001	0.0002	0.0002	0.0001	0.0002	0.0004
208×60 (2)		15	1	1	56	26	23	14
Spect	74.91/86.14	78.65/82.40	74.16/79.40	71.91/79.40	74.91/86.14	80.15/83.52	79.03/79.90	**82.77/86.90**
		0.0001	0.0001	0.0002	0.0002	0.0001	0.0002	0.0002
267×44 (2)		5	1	1	44	4	3	3
WPBC	63.64/60.10	64.14/81.31	57.58/76.26	78.28/76.26	74.24/76.77	63.64/67.68	64.14/81.31	**80.81/86.26**
		0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
198×34 (2)		2	1	1	8	3	2	1
Dermatology	90.16/98.09	92.35/**98.36**	72.68/97.27	96.72/97.27	96.72/**98.36**	92.90/**98.36**	89.62/**98.36**	**97.54/98.36**
		0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
366×34 (6)		11	2	5	31	6	33	11
Ecoli	81.25/91.96	81.25/**91.96**	80.36/84.52	82.14/85.12	81.25/**91.96**	81.85/**91.96**	81.25/**91.96**	**82.44/91.96**
		0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
336×7 (8)		7	5	5	7	6	7	7
Seeds	90.48/99.05	90.95/98.10	88.10/94.76	97.14/95.71	90.48/**99.05**	90.48/98.10	90.95/98.10	**98.10/99.05**
		0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
210×7 (3)		4	2	2	7	6	4	4
Wine	76.97/91.19	76.97/93.26	70.79/93.15	98.31/94.04	96.63/96.07	76.97/95.51	76.97/94.38	**100.0/98.88**
		0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
178×13 (3)		6	1	2	8	8	6	7
Zoo	98.02/96.04	98.02/96.25	97.03/97.13	98.02/95.15	98.02/96.04	**99.01**/96.04	98.02/**97.03**	98.02/**97.03**
		0.0001	0.0001	0.0014	0.0001	0.0001	0.0001	0.0001
101×16 (7)		2	4	4	16	6	13	13
Mean	78.05/86.49	79.10/88.32	73.79/86.01	85.51/87.18	86.05/90.19	80.37/88.77	79.20/89.14	**90.55/92.89**

is a diagonal matrix and $M(i,i) = \sum_{j=1}^{k} \frac{m_j}{m-m_j} - \frac{m_{y_i}}{m-m_{y_i}}$ with $i = 1, 2, \ldots, m$.

Combined (17) with (18), we have

$$\begin{aligned} w^\top S_w^\phi w &= a^\top K(X,X)DK(X,X)a \\ w^\top S_b^{*\phi} w &= a^\top K(X,X)(E - \gamma_1 E' + \gamma_2 M)K(X,X)a. \end{aligned} \quad (19)$$

Thus, our kernel based RDA maximizes

$$J_{RDA}^\phi(w) = \frac{w^\top S_b^{*\phi} w}{w^\top S_w^\phi w}, \quad (20)$$

i.e.,

$$J_{RDA}^\phi(a) = \frac{a^\top K(X,X)(E - \gamma_1 E' + \gamma_2 M)K(X,X)a}{a^\top K(X,X)DK(X,X)a}. \quad (21)$$

The solution to (21) can be obtained by solving following generalized eigenvalue problem

$$(E - \gamma_1 E' + \gamma_2 M)K(X,X)a = \lambda DK(X,X)a. \quad (22)$$

After solving the problem (22), we can obtain m generalized eigenvectors a. Then, some eigenvectors can be selected by their corresponding eigenvalues similar to linear case. Though we cannot get the explicit mapping vector w, the mapped sample x can be obtained explicitly by (17).

IV. EXPERIMENTS

In this section, we analyze the performance of our RDA compared with PCA and LDA with its extensions. The dimensionality reduction methods, including PCA [1], PCA+LDA [3], LDA [8], OLSDA [11], OCM [13], MMC [14], and our RDA were implemented by Matlab [32] on a PC with an Intel

TABLE 2. **LOO results by NN/SVM classifier with kernel based dimensionality reduction on the benchmark datasets.**

	Baseline	PCA	PCA+LDA	LDA	OLSDA	OCM	MMC	RDA
Data m×n (k)	Acc./Acc.	Acc./Acc. Time(Sec.) Dim.	Acc./Acc. Time(Sec.) Dim.	Acc./Acc. Time(Sec.) Dim.	Acc./Acc. Time(Sec.) Dim.	Acc./Acc. Time(Sec.) Dim.	Acc./Acc. Time(Sec.) Dim.	Acc./Acc. Time(Sec.) Dim.
Heartc 303×14 (2)	84.82/91.71	84.49/**94.72** 0.0001 7	92.41/94.13 0.0066 1	99.01/93.80 0.0442 1	**99.67**/93.80 0.0035 9	91.09/94.06 0.0030 10	85.81/93.73 0.0037 7	**99.67**/93.13 0.0461 2
Heartstatlog 270×13 (2)	81.11/84.48	78.89/86.30 0.0001 9	**82.22**/82.22 0.0029 1	70.74/85.93 0.0278 1	78.15/90.00 0.0020 3	79.26/86.67 0.0017 13	78.15/84.44 0.0019 10	74.81/**90.37** 0.0363 6
Hepatitis 155×19 (2)	80.00/80.00	80.65/81.74 0.0001 5	85.16/85.35 0.0010 1	83.23/80.94 0.0055 1	86.45/80.00 0.0007 11	81.29/83.35 0.0007 2	81.29/83.35 0.0007 19	**90.32/89.29** 0.0077 7
Hourse 300×26 (2)	77.33/74.67	76.33/73.67 0.0001 3	76.33/73.67 0.0049 1	70.67/**92.00** 0.0042 1	82.00/74.67 0.0026 24	73.00/75.00 0.0026 6	73.67/75.00 0.0029 8	**91.00**/91.67 0.0045 1
Housevotes 435×16 (2)	88.74/90.42	**92.64**/92.16 0.0001 15	**92.64**/87.82 0.0001 1	85.29/90.29 0.1143 1	90.80/91.38 0.0079 16	92.18/**92.87** 0.0075 4	**92.64**/92.09 0.0075 15	91.49/92.18 0.1224 2
Ionosphere 351×33 (2)	72.93/84.30	94.87/93.46 0.0001 8	91.45/93.24 0.0001 1	79.20/94.11 0.0594 1	94.59/**94.20** 0.0054 15	**95.44**/90.03 0.0042 32	94.87/86.04 0.0055 6	90.31/**94.20** 0.0680 23
Sonar 208×60 (2)	72.12/72.60	85.58/85.58 0.0001 17	83.65/86.12 0.0001 1	80.77/86.88 0.0150 1	89.42/86.60 0.0017 9	83.65/86.60 0.0012 48	85.58/85.08 0.0016 29	**95.67/91.44** 0.0165 1
Spect 267×44 (2)	80.52/70.04	80.52/77.15 0.0001 26	79.40/76.67 0.0027 1	79.78/79.73 0.0268 1	81.27/80.04 0.0022 37	81.27/80.79 0.0022 42	80.15/80.16 0.0024 11	**97.75/81.16** 0.0329 2
WPBC 198×34 (2)	76.26/70.71	78.79/70.27 0.0001 3	76.26/75.69 0.0016 1	82.32/76.26 0.0153 1	78.79/**79.90** 0.0011 1	76.26/70.71 0.0011 9	87.88/75.76 0.0011 1	**100.0/79.90** 0.0153 1
Dermatology 366×34 (6)	91.80/95.09	96.99/97.27 0.0001 7	96.45/98.09 0.0035 5	92.90/93.07 0.0712 4	96.17/**98.63** 0.0068 33	96.45/98.36 0.0048 9	**97.27**/97.27 0.0073 7	95.90/98.09 0.0896 6
Ecoli 336×7 (8)	79.46/83.04	**83.33/84.23** 0.0001 5	81.25/81.48 0.0001 5	62.50/80.56 0.0540 7	71.43/81.86 0.0037 7	**83.33**/83.76 0.0036 7	82.44/82.65 0.0037 5	69.64/83.56 0.0544 7
Seeds 210×7 (3)	92.86/90.19	92.38/93.33 0.0001 5	90.48/92.86 0.0001 2	83.81/80.48 0.0185 2	91.90/84.76 0.0016 6	91.90/92.86 0.0013 5	92.86/93.33 0.0015 7	**93.33/94.05** 0.0148 7
Wine 178×13 (3)	98.31/98.31	97.75/**98.31** 0.0001 8	**98.31**/93.82 0.0011 2	88.20/97.19 0.0083 2	**98.31**/96.63 0.0011 3	97.75/96.07 0.0008 5	**98.31/98.31** 0.0011 4	97.75/**98.31** 0.0088 12
Zoo 101×16 (7)	92.08/93.03	95.05/92.18 0.0001 3	95.05/94.26 0.0001 2	90.10/93.56 0.0018 5	**98.02**/92.08 0.0004 13	97.03/95.05 0.0003 6	96.04/97.13 0.0004 3	96.04/**97.59** 0.0026 13
Mean	83.45/90.29	87.01/87.16	87.22/86.81	82.03/87.48	88.35/87.46	87.13/87.58	87.64/87.45	**91.69/91.06**

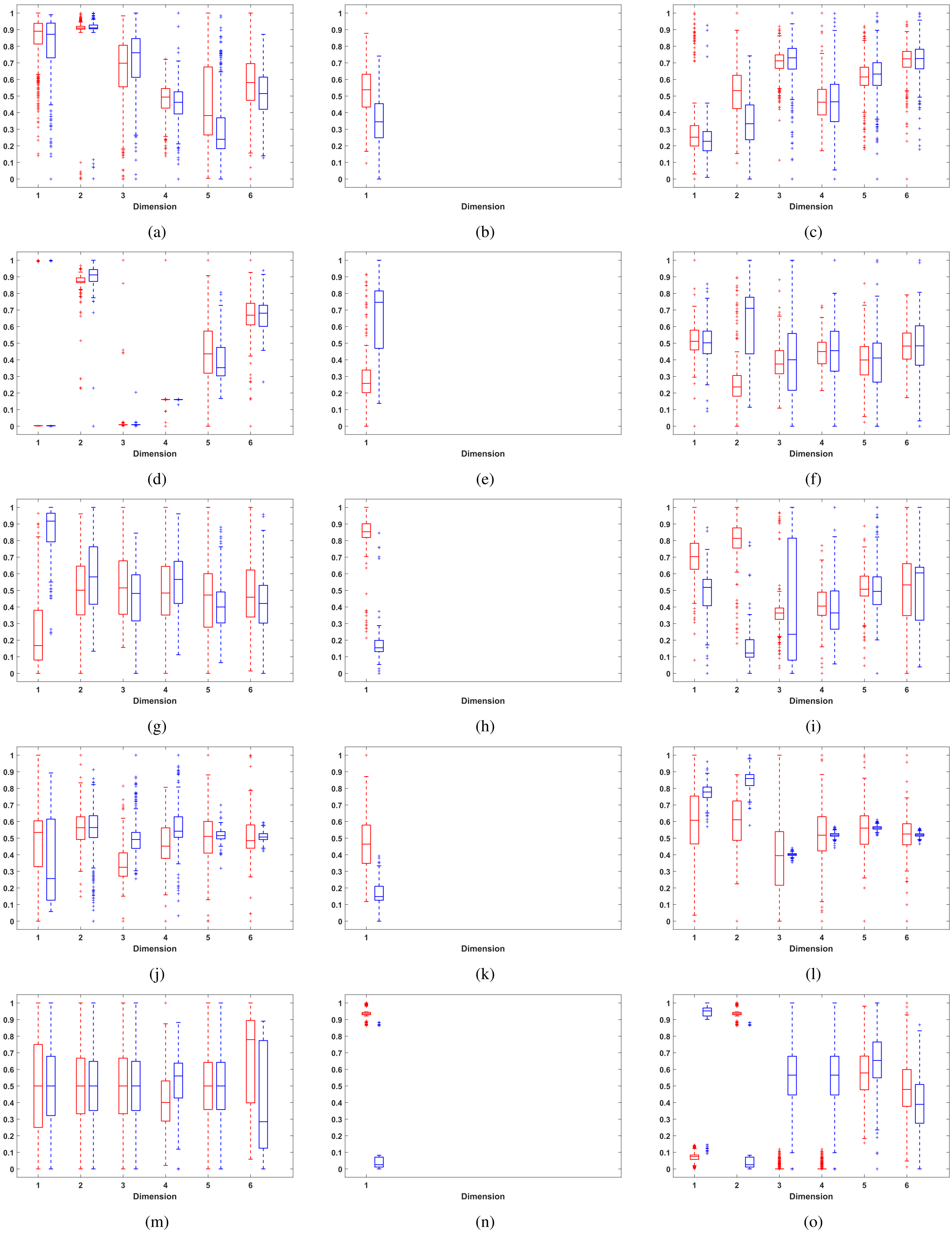
TABLE 3. LOO results by NN/SVM classifier with linear dimensionality reduction on the benchmark datasets.

	Baseline	PCA	PCA+LDA	LDA	OLSDA	OCM	MMC	RDA
Data m×n (k)	Acc./Acc.	Acc./Acc. Time(Sec.) Dim.	Acc./Acc. Time(Sec.) Dim.	Acc./Acc. Time(Sec.) Dim.	Acc./Acc. Time(Sec.) Dim.	Acc./Acc. Time(Sec.) Dim.	Acc./Acc. Time(Sec.) Dim.	Acc./Acc. Time(Sec.) Dim.
Australian 690×14 (2)	65.51/94.94	68.12/94.35 0.0001 5	61.74/94.49 0.0005 1	80.29/94.49 0.0065 1	80.29/87.39 0.0014 7	66.52/94.35 0.0014 7	66.52/93.77 0.0008 6	**84.35/95.65** 0.0054 9
German 1000×20 (2)	61.20/70.60	61.20/73.20 0.0001 16	60.10/70.00 0.0002 1	69.20/70.30 0.0006 1	69.70/70.00 0.0004 17	61.20/69.70 0.0004 19	61.20/70.10 0.0004 16	**73.50/74.30** 0.0005 15
Tictactoe 958×27 (2)	67.43/98.06	98.75/**98.83** 0.0001 7	67.33/97.06 0.0005 1	95.20/98.33 0.0011 1	98.33/98.33 0.0007 14	97.18/97.41 0.0005 1	**99.16/98.83** 0.0006 15	98.43/98.43 0.0016 12
TIC 5822×85 (2)	89.14/91.02	89.85/93.85 0.0001 4	89.80/90.07 0.0011 1	89.80/94.01 0.0040 1	89.92/**94.02** 0.0033 65	89.78/92.69 0.0023 1	89.59/93.66 0.0032 7	**90.07/94.02** 0.0048 44
Two 7400×20 (2)	94.68/97.77	96.61/97.80 0.0001 3	96.50/97.80 0.0007 1	96.57/97.78 0.0007 1	94.96/97.64 0.0005 17	**97.07**/97.84 0.0005 1	96.78/**97.85** 0.0005 1	96.95/97.80 0.0011 3
Car 1728×6 (4)	87.09/86.97	87.85/90.97 0.0001 6	81.71/82.02 0.0003 2	79.17/80.02 0.0002 2	95.20/95.97 0.0002 3	93.98/95.85 0.0001 5	88.31/93.73 0.0001 4	**96.64/96.16** 0.0002 6
Letters 20000×16 (26)	96.25/98.58	96.25/**98.58** 0.0001 16	95.47/96.20 0.0009 11	95.94/96.44 0.0009 15	**97.52**/98.19 0.0005 15	96.84/96.99 0.0004 14	96.22/**98.58** 0.0005 16	96.21/**98.58** 0.0017 15
Vehicle 846×18 (4)	65.25/92.08	65.25/89.95 0.0001 15	42.67/75.46 0.0001 1	73.88/92.08 0.0006 3	72.70/92.20 0.0038 12	65.25/92.32 0.0003 17	65.25/89.83 0.0004 16	**79.67/92.70** 0.0005 6
Mean	78.32/91.25	82.99/92.19	74.42/87.88	85.01/90.43	87.33/91.71	83.48/92.14	82.88/92.04	**89.48/93.45**

Core Duo processor (4.2 GHz) with 16 GB RAM. The nearest neighbor (NN) classifier [33] with Euclidean metric and support vector machines (SVM) are used for classification in the experiments, where the parameter c in SVM is selected from $\{2^i | i = -8, -7, \ldots, 7\}$. The parameters γ_1 and γ_2 are selected from $\{2^i | i = -8, -7, \ldots, 7\}$. For kernel based methods, the Gaussian kernel $K(x, y) = \exp\{-\mu||x - y||^2\}$ [34] is used and its parameter μ is selected from $\{2^i | i = -10, -7, \ldots, 5\}$.

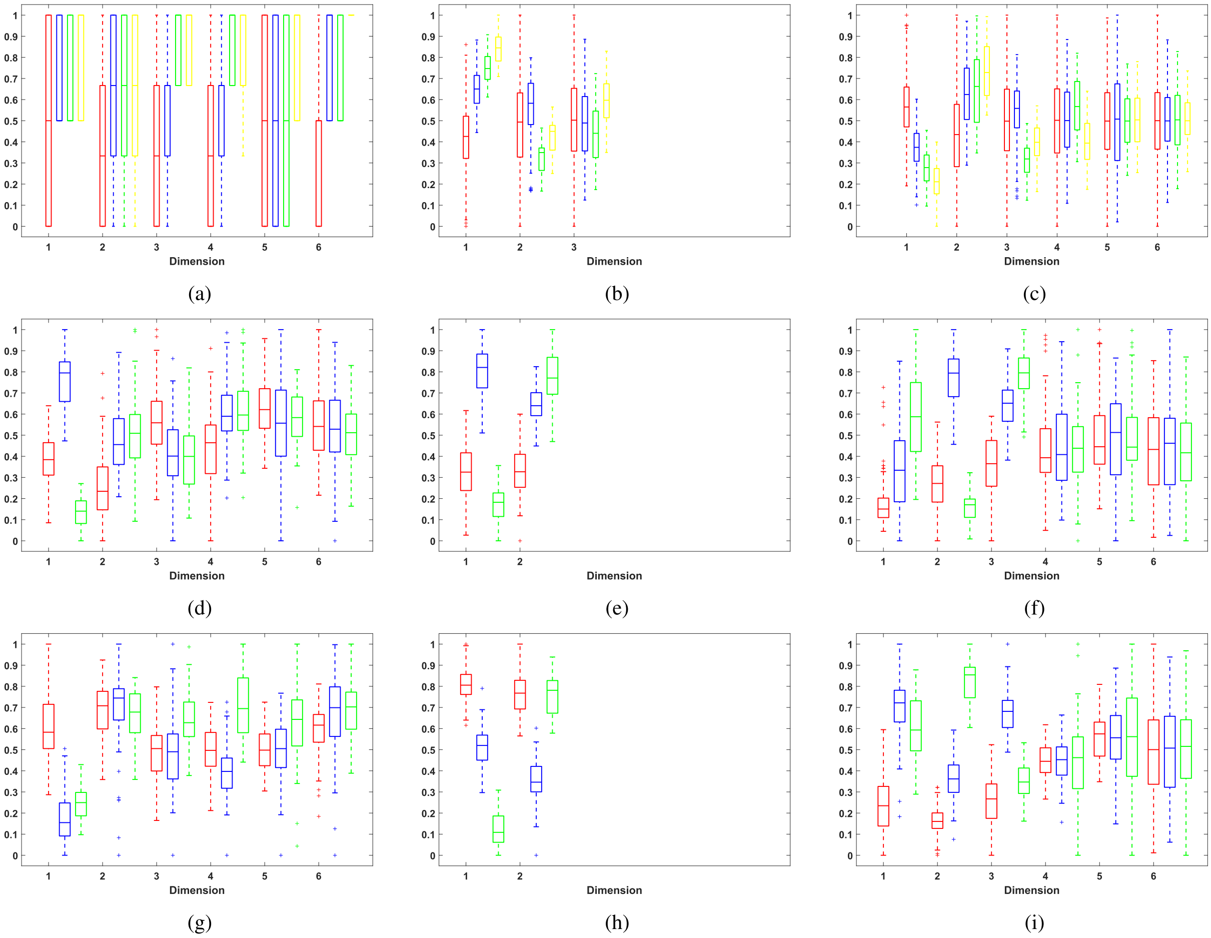
First, we test these methods on some benchmark datasets [35]. To clearly show their best performances, the accuracy was calculated for all possible dimensions by the leave-one-out (LOO) technique [36], [37], and the highest accuracy (%) with its corresponding dimension (No.) was recorded. Table 1 shows the results on 14 datasets where the sample number of each dataset is no more than 500, and Table 3 shows the results on 8 datasets with the sample number over 500. The highest accuracies are bolded and the mean accuracies are also computed. From Tables 1 and 3, it is clear that RDA exhibits a much better performance on most datasets than other methods. There are many datasets on which RDA works much better than others, and there are few datasets on which RDA works much worse than others. Moreover, the features obtained by our RDA are more useful than other methods. For example, RDA owns the highest accuracies on "Hepatitis" and "WPBC" with only one feature. And the second-best OLSDA needs equal or more features than RDA on 17/22 datasets. Table 2 shows the results for kernel based dimensionality reduction, and RDA also has the highest mean accuracy than other methods. Due to all these dimensionality reduction methods solved an eigenvalue or generalized eigenvalue problem, we also reported the main computation time in Tables 1, 2 and 3. From these tables, it is obvious that the difference of the main computation costs is within 1 second. Therefore, there is no significant difference on the learning speed among these methods.

To further exhibit the performance of our RDA compared with PCA and LDA, the dimensional reduction results are depicted for each feature on 8 datasets (we just depict the box plot for the first 6 features to have a clear comparison). Figs. 3 and 4 shows the box plots of these methods on binary and multiple class datasets, respectively. The horizontal axis denotes the feature sequence number, and different classes are depicted with different colors. The samples for each feature are normalized to [0, 1]. From Figs. 3 and 4, we observed that: (i) the samples from different classes cannot be separated well by PCA for each feature; (ii) LDA generally can obtain a well separation of two classes on binary class datasets with its only feature; (iii) RDA always owns a well separation by one feature similar to LDA, and the overlapped samples by this feature can often be separated by other features. Thus, our RDA has a better classification performance than PCA and LDA. For example, on the "Ionosphere" dataset, our RDA obtains 13% higher accuracy than LDA by adding additional 3 features; and on the "Car" dataset, our RDA obtains 17% higher accuracy than LDA by adding additional 4 features. In fact, RDA performs better than PCA on 20/22 datasets, and does better than LDA on 21/22 datasets, from Tables 1 and 3. Therefore, our RDA can obtain much more useful features than PCA and LDA on these datasets.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l) (m) (n) (o)

FIGURE 3. Dimensionality reduction results for the first 6 features on the binary class datasets, where the subtitle of each figure denotes the dataset and the method used, e.g., (a) German(PCA) shows the results of PCA on the "German" dataset. For each figure, there are two classes depicted by two colored boxes, where the horizontal axis denotes the feature sequence number. The box shows the distribution of the samples for one dimension, which includes the minimum, lower quartile, median, upper quartile, maximum, and some outliers. (a) German(PCA). (b) German(LDA). (c) German(RDA). (d) Hourse(PCA). (e) Hourse(LDA). (f) Hourse(RDA). (g) Housevotes(PCA). (h) Housevotes(LDA). (i) Housevotes(RDA). (j) Ionosphere(PCA). (k) Ionosphere(LDA). (l) Ionosphere(RDA). (m) Tictactoe(PCA). (n) Tictactoe(LDA). (o) Tictactoe(RDA).



(a) (b) (c)

(d) (e) (f)

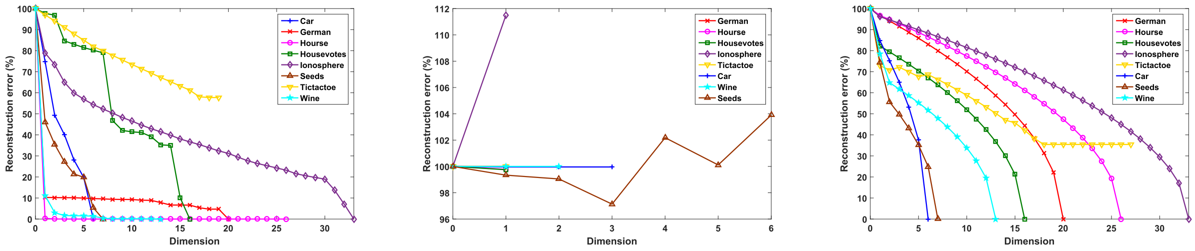
(g) (h) (i)

FIGURE 4. Dimensionality reduction results for the first 6 features on the multi-class datasets, where the subtitle of each figure denotes the dataset and the method used. For each feature, there are multiple classes depicted by different colored boxes. (a) Car(PCA). (b) Car(LDA). (c) Car(RDA). (d) Seeds(PCA). (e) Seeds(LDA). (f) Seeds(RDA). (g) Wine(PCA). (h) Wine(LDA). (i) Wine(RDA).

TABLE 4. Classification results by the dimensionality reduction methods on the YALE and ORL datasets.

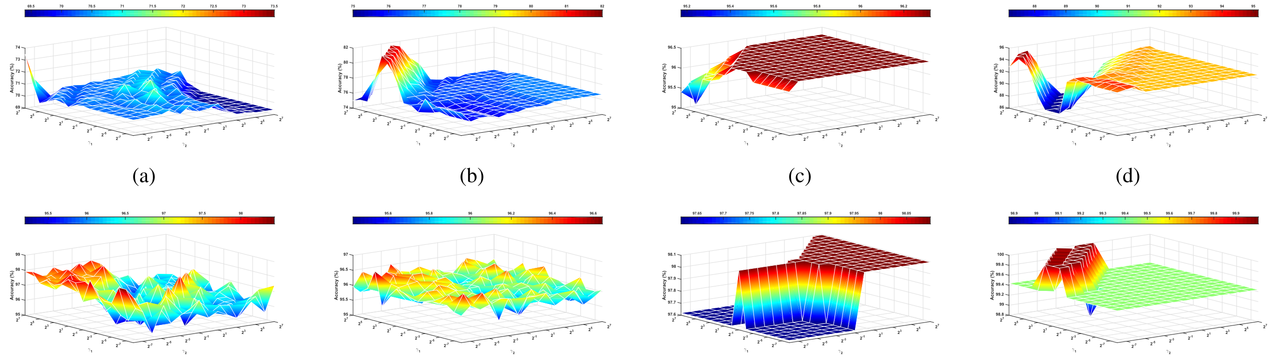
Data	Baseline	PCA	PCA+LDA	LDA	OLSDA	OCM	MMC	RDA
YALE								
2	43.44±3.94	41.24±3.63	51.84±4.78	51.84±4.78	58.61±4.92	47.07±4.52	43.44±3.94	**59.62**±4.92
3	49.42±4.20	47.95±4.30	64.53±4.85	70.53±3.93	**71.55**±3.67	55.10±4.34	49.42±4.20	71.43±3.60
4	52.65±3.97	51.81±4.17	71.33±4.40	76.93±4.29	78.48±4.17	58.59±3.61	52.69±4.00	**79.23**±4.10
5	56.22±4.13	55.58±4.17	77.22±3.84	82.16±3.66	82.53±3.12	62.60±4.45	56.22±4.13	**84.54**±3.28
6	58.69±4.73	59.17±4.64	81.20±3.74	84.29±4.31	85.04±3.81	65.60±4.51	58.93±5.03	**86.05**±4.11
7	60.20±4.90	60.53±5.52	83.10±4.14	86.00±3.67	86.87±3.17	67.33±4.74	60.20±4.89	**87.77**±3.39
8	63.64±5.05	64.09±5.52	85.20±4.39	89.07±3.64	89.87±3.51	70.31±5.42	64.44±5.85	**90.84**±3.58
ORL								
2	66.92±3.47	69.05±3.31	**71.64**±3.45	65.42±5.29	71.69±3.34	69.89±1.97	68.91±3.15	71.45±3.35
3	76.64±2.34	77.54±2.64	79.63±1.94	66.04±14.01	79.53±2.90	77.12±2.16	77.23±2.72	**79.66**±2.76
4	82.14±2.22	83.55±1.99	83.68±1.88	69.79±12.65	83.04±2.29	79.92±2.06	83.08±2.06	**85.35**±2.03
5	86.35±2.41	87.18±2.12	86.13±1.48	73.57±11.29	85.93±2.60	81.90±2.11	86.85±2.28	**88.63**±2.27
6	88.57±2.27	89.65±1.96	88.98±1.51	74.06±11.27	86.34±2.99	83.53±2.98	89.43±2.34	**91.30**±1.97
7	91.32±2.39	92.10±2.60	90.25±1.45	74.10±10.48	87.88±2.59	84.33±2.62	92.02±2.54	**93.35**±2.45
8	92.55±2.41	93.95±2.58	94.05±1.52	70.45±18.10	89.45±2.84	85.81±3.07	94.10±2.44	**94.90**±2.45

In the following, we test the reconstruction of RDA compared with PCA and LDA. These methods were implemented on the above 8 benchmark datasets, and the reconstruction error (%) with the increasing features was depicted in Fig. 5. It is obvious that LDA cannot recover the primal space, while PCA and RDA can recover that space on most datasets except "Tictactoe". On the dataset "Tictactoe", though PCA and RDA cannot recover the primal space, our RDA can recover 20% more information than PCA. Moreover, the reconstruction error curves of our RDA are more uniform than PCA



(a) (b) (c)

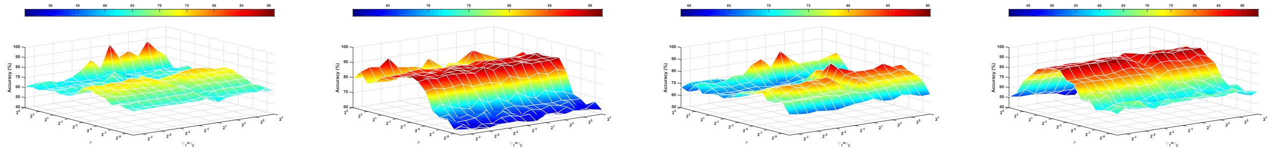
FIGURE 5. Reconstruction error (%) with the increasing features for PCA, LDA, and RDA. (a) PCA. (b) LDA. (c) RDA.



(a) (b) (c) (d)

(e) (f) (g) (h)

FIGURE 6. Influence of parameters in RDA for linear case. (a) German. (b) Hourse. (c) Housevotes. (d) Ionosphere. (e) Tictactoe. (f) Car. (g) Seeds. (h) Wine.

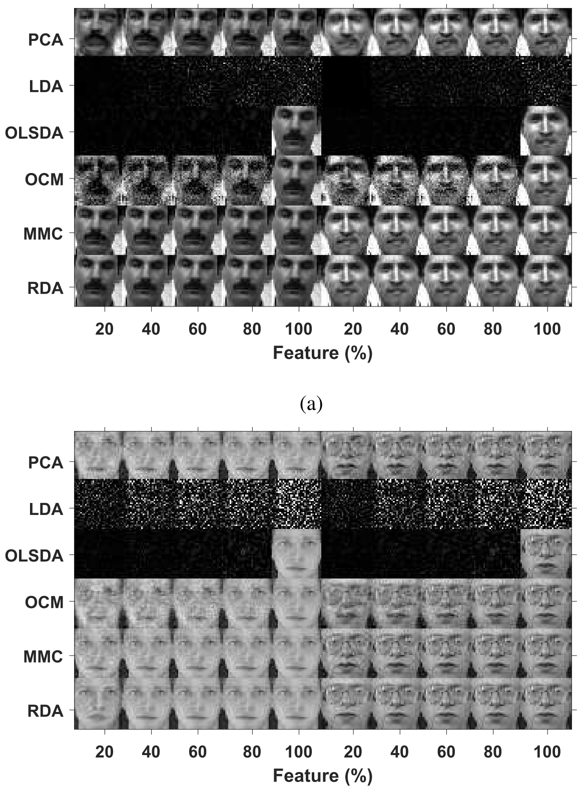


(a) (b) (c) (d)

FIGURE 7. Influence of parameters in kernel based RDA. (a) Hourse. (b) Housevotes. (c) Ionosphere. (d) Seeds.

on these datasets. Thus, the primal space would be recovered by RDA more homogenously than PCA. Fig. 6 shows the influence of parameters in RDA on the 8 datasets, and Fig. 7 shows the influence in kernel based RDA. From Figs. 6 and 7, we observed that the parameters γ_1 and γ_2 significantly affect the performance of linear RDA, while in kernel based RDA the parameter μ plays an more important role compared with $\gamma_1 = \gamma_2$.

To further evaluate the performance of RDA, we experimented these methods on the image dimensionality reduction (including YALE and ORL face datasets [38], [39]) by NN classifier. The YALE dataset contains 165 images of 15 individuals under various facial expressions, where each individual has 11 different images. The YALE dataset was grouped into two parts the same as in [38]. One part is used for training and the other is used for testing. In this experiment, the number of training images chosen for each individual is 2, 3, 4, 5, 6, 7, and 8, respectively, from which we obtain seven training subsets. The ORL dataset contains 400 images of 40 individuals, and it was also grouped the same as the YALE dataset. The image size of YALE and ORL is uniformed to 32 × 32. The dimension of the testing dataset was reduced by the mapping which was learned from the training datasets (where the reduced dimension is fixed to 50), and then the NN classifier was employed to predict the testing samples. Table 4 shows the average accuracies and the standard deviations of these methods on the YALE and ORL datasets. The bold number in Table 4 highlights the highest classification accuracy on each training subset. From Table 4 we see that, on most cases, these dimensionality reduction methods greatly improve the performance of NN classifier, and the accuracy increases with the training set for each method. Thereinto, our RDA performs much better than other methods. To further exhibit the reconstruction ability



(a)

(b)

FIGURE 8. Reconstructed faces by the dimensionality reduction methods from 20%-100% features on the YALE and ORL faces. (a) YALE. (b) ORL.

of these methods, we depicted the reconstructed faces from $20\% - 100\%$ features in Fig. 8. It can be seen from Fig. 8 that our RDA owns the best reconstruction performance among these methods. LDA and OLSDA cannot reconstruct the faces, unless using the whole features in OLSDA. Not only our RDA can recover the faces better with more features, but the recovered faces by RDA is much better than PCA, OCM, and MMC with the same features.

V. CONCLUSION

In this paper, a linear dimensionality reduction method based on a new defined between-class scatter has been proposed, called RDA. Since the new between-class scatter matrix is generally full-rank, RDA obtains a full-rank mapping matrix. Therefore, RDA reduces the sample space to arbitrary dimension and the mapped samples can be recovered. More dimensionality features greatly improve the performance of RDA, and the primal space can be recovered by the whole features. Preliminary experiments on several benchmark datasets confirm the better performance of RDA compared with other dimensionality reduction methods. For convenience, the Matlab code of our RDA is uploaded upon http://www.optimal-group.org/Resources/Code/RDA.html. Due to RDA hires a similar formation as LDA, the future work includes extending RDA into other between-class scatter based methods, such as L1-norm dimensionality reduction [10], [40] and 2D dimensionality reduction [41], [42].

REFERENCES

[1] B. C. Moore, “Principal component analysis in linear systems: Controllability, observability, and model reduction,” *IEEE Trans. Autom. Control*, vol. AC-26, no. 1, pp. 17–32, Feb. 1981.

[2] I. Jolliffe, *Principal Component Analysis*. Hoboken, NJ, USA: Wiley, 2005.

[3] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman, “Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[4] J. Yang and J.-Y. Yang, “Why can LDA be performed in PCA transformed space?” *Pattern Recognit.*, vol. 36, no. 2, pp. 563–566, 2003.

[5] L. I. Kuncheva and W. J. Faithfull, “PCA feature extraction for change detection in multidimensional unlabeled data,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 1, pp. 69–80, Jan. 2014.

[6] G.-F. Lu, Y. Wang, J. Zou, and Z. Wang, “Matrix exponential based discriminant locality preserving projections for feature extraction,” *Neural Netw.*, vol. 97, pp. 127–136, Jan. 2018.

[7] K. W. Jorgensen and L. K. Hansen, “Model selection for Gaussian kernel PCA denoising,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 1, pp. 163–168, Jan. 2012.

[8] S. Balakrishnama and A. Ganapathiraju, “Linear discriminant analysis—A brief tutorial,” *Inst. Signal Inf. Process.*, vol. 18, pp. 1–8, Mar. 1998.

[9] C. H. Park and H. Park, “A comparison of generalized linear discriminant analysis algorithms,” *Pattern Recognit.*, vol. 41, no. 3, pp. 1083–1097, 2008.

[10] H. Wang, X. Lu, Z. Hu, and W. Zheng, “Fisher discriminant analysis with L1-norm,” *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 828–842, Jun. 2014.

[11] F. Nie, S. Xiang, Y. Liu, C. Hou, and C. Zhang, “Orthogonal vs. Uncorrelated least squares discriminant analysis for feature extraction,” *Pattern Recognit. Lett.*, vol. 33, no. 5, pp. 485–491, Apr. 2012.

[12] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, and G. Zhang, “Transfer learning using computational intelligence: A survey,” *Knowl.-Based Syst.*, vol. 80, pp. 14–23, May 2015.

[13] H. Park, M. Jeon, and J. Ben Rosen, “Lower dimensional representation of text data based on centroids and least squares,” *BIT Numer. Math.*, vol. 43, no. 2, pp. 427–448, 2003.

[14] H. R. Li, T. Jiang, and K. Zhang, “Efficient and robust feature extraction by maximum margin criterion,” *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 157–165, Feb. 2006.

[15] B. Xu, K. Huang, and C.-L. Liu, “Dimensionality reduction by minimal distance maximization,” in *Proc. 20th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2010, pp. 569–572.

[16] Y. Zhang and D.-Y. Yeung, “Worst-case linear discriminant analysis,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 2568–2576.

[17] L. Yang and S. Chen, “Linear discriminant analysis with worst between-class separation and average within-class compactness,” *Frontiers Comput. Sci.*, vol. 8, no. 5, pp. 785–792, 2014.

[18] Z. Wang, Y.-H. Shao, L. Bai, C.-N. Li, L.-M. Liu, and N.-Y. Deng, “MBLDA: A novel multiple between-class linear discriminant analysis,” *Inf. Sci.*, vol. 369, pp. 199–220, Nov. 2016.

[19] O. L. Mangasarian and E. W. Wild, “Multisurface proximal support vector machine classification via generalized Eigenvalues,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 69–74, Jan. 2006.

[20] Y. H. Shao, C. H. Zhang, X. B. Wang, and N. Y. Deng, “Improvements on twin support vector machines,” *IEEE Trans. Neural Netw.*, vol. 22, no. 6, pp. 962–968, May 2011.

[21] Z. Wang, Y.-H. Shao, L. Bai, C.-N. Li, L.-M. Liu, and N.-Y. Deng, “Insensitive stochastic gradient twin support vector machines for large scale problems,” *Inf. Sci.*, vol. 462, pp. 114–131, Sep. 2018.

[22] M. Loog, R. P. W. Duin, and R. Haeb-Umbach, “Multiclass linear dimension reduction by weighted pairwise Fisher criteria,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 7, pp. 762–766, Jul. 2001.

[23] Y. Wang, Y. Jiang, Y. Wu, and Z.-H. Zhou, “Multi-manifold clustering,” in *Trends in Artificial Intelligence*. Berlin, Germany: Springer, 2010, pp. 280–291.

[24] F.-X. Song, K. Cheng, J.-Y. Yang, and S.-H. Liu, “Maximum scatter difference, large margin linear projection and support vector machines,” *Acta Autom. Sin.*, vol. 30, no. 6, pp. 890–896, 2004.

[25] L. Chen, H. Liao, M. Ko, J. Lin, and G. Yu, "A new LDA-based face recognition system which can solve the small sample size problem," *Pattern Recognit.*, vol. 33, no. 10, pp. 1713–1726, 2000.

[26] H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional data—With application to face recognition," *Pattern Recognit.*, vol. 34, no. 10, pp. 2067–2070, 2001.

[27] W. Zheng, L. Zhao, and C. Zou, "An efficient algorithm to solve the small sample size problem for LDA," *Pattern Recognit.*, vol. 37, no. 5, pp. 1077–1079, 2004.

[28] J. Shao, Y. Wang, X. Deng, and S. Wang, "Sparse linear discriminant analysis by thresholding for high dimensional data," *Ann. Statist.*, vol. 39, no. 2, pp. 1241–1265, 2011.

[29] J. H. Friedman, "Regularized discriminant analysis," *J. Amer. Statist. Assoc.*, vol. 84, no. 405, pp. 165–175, 1989.

[30] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Mullers, "Fisher discriminant analysis with kernels," in *Proc. IEEE Signal Process. Soc. Workshop Neural Netw. Signal Process.*, Aug. 1999, pp. 41–48.

[31] C. H. Park and H. Park, "Nonlinear discriminant analysis using kernel functions and the generalized singular value decomposition," *SIAM J. Matrix Anal. Appl.*, vol. 27, no. 1, pp. 87–102, 2005.

[32] The MathWorks, Inc. (2018). *MATLAB., User's Guide*. [Online]. Available: http://www.mathworks.com

[33] D. T. Larose, "k-nearest neighbor algorithm," *Discovering Knowledge in Data: An Introduction to Data Mining*. Hoboken, NJ, USA: Wiley, 2005, pp. 90–106.

[34] D. Kakde, A. Chaudhuri, S. Kong, M. Jahja, H. Jiang, and J. Silva, "Peak criterion for choosing Gaussian kernel bandwidth in support vector data description," in *Proc. IEEE Int. Conf. Prognostics Health Manage. (ICPHM)*, Jun. 2017, pp. 32–39.

[35] C. Blake and C. Merz. (2014). *UCI Repository for Machine Learning Databases*. [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

[36] A. Vehtari, A. Gelman, and J. Gabry, "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC," *Statist. Comput.*, vol. 27, no. 5, pp. 1413–1432, 2017.

[37] T. F. Y. Vicente, M. Hoai, and D. Samaras, "Leave-one-out kernel optimization for shadow detection and removal," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 682–695, Mar. 2018.

[38] D. Cai, X. He, Y. Hu, J. Han, and T. Huang. (2007). *Learning a Spatially Smooth Subspace for Face Recognition*. [Online]. Available: http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html

[39] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using Laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.

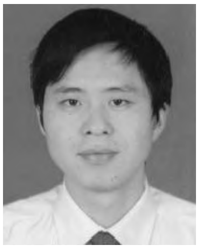
[40] C.-N. Li, Z.-R. Zheng, M.-Z. Liu, Y.-H. Shao, and W.-J. Chen, "Robust recursive absolute value inequalities discriminant analysis with sparseness," *Neural Netw.*, vol. 93, pp. 205–218, Sep. 2017.

[41] M. Li and B. Yuan, "2D-LDA: A statistical linear discriminant analysis for image matrix," *Pattern Recognit. Lett.*, vol. 26, no. 5, pp. 527–532, 2005.

[42] C.-N. Li, Y.-H. Shao, and N.-Y. Deng, "Robust L1-norm two-dimensional linear discriminant analysis," *Neural Netw.*, vol. 65, pp. 92–104, May 2015.

LAN BAI received the Ph.D degree from the Department of Mathematics from Jilin University, China, in 2014. She is currently a Lecturer with the School of Mathematics, Inner Mongolia University. Her research interests include data mining, machine learning, and optimization methods.

ZHEN WANG received the bachelor's, master's, and Ph.D. degrees from the College of Mathematics, Jilin University, China, in 2006, 2010, and 2014, respectively. He is currently an Associate Professor with the School of Mathematical Sciences from Inner Mongolia University. His research interests include pattern recognition, text categorization, and data mining.



YUAN-HAI SHAO received the bachelor's degree in information and computing science from the College of Mathematics, Jilin University, in 2006, and the master's degree in applied mathematics and the Ph.D. degree in operations research and management from the College of Science, China Agricultural University, China, in 2008 and 2011, respectively. He is currently a Professor with the School of Economics and Management, Hainan University. His research interests include data mining, machine learning, and optimization methods. He has published over 80 refereed papers on these areas.

CHUN-NA LI received the master's and Ph.D. degrees from the Department of Mathematics, Harbin Institute of Technology, China, in 2009 and 2012, respectively. She is currently an Associate Professor with the Zhijiang College, Zhejiang University of Technology. Her research interests include data mining, machine learning, and optimization methods.