Hindawi Publishing Corporation
Computational Intelligence and Neuroscience
Volume 2016, Article ID 4920670, 12 pages
http://dx.doi.org/10.1155/2016/4920670

Research Article

Regularized Embedded Multiple Kernel Dimensionality Reduction for Mine Signal Processing

Shuang Li,[1] Bing Liu,[2] and Chen Zhang[2]

[1]*School of Management, China University of Mining and Technology, Xuzhou, Jiangsu 221116, China*
[2]*School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, Jiangsu 221116, China*

Correspondence should be addressed to Bing Liu; liubing@cumt.edu.cn and Chen Zhang; zc@cumt.edu.cn

Received 22 November 2015; Revised 4 March 2016; Accepted 11 April 2016

Academic Editor: Hong Man

Copyright © 2016 Shuang Li et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Traditional multiple kernel dimensionality reduction models are generally based on graph embedding and manifold assumption. But such assumption might be invalid for some high-dimensional or sparse data due to the curse of dimensionality, which has a negative influence on the performance of multiple kernel learning. In addition, some models might be ill-posed if the rank of matrices in their objective functions was not high enough. To address these issues, we extend the traditional graph embedding framework and propose a novel regularized embedded multiple kernel dimensionality reduction method. Different from the conventional convex relaxation technique, the proposed algorithm directly takes advantage of a binary search and an alternative optimization scheme to obtain optimal solutions efficiently. The experimental results demonstrate the effectiveness of the proposed method for supervised, unsupervised, and semisupervised scenarios.

1. Introduction

Dimensionality reduction (DR) methods in supervised, unsupervised, and semisupervised learning tasks have attracted much attention in computer vision and pattern recognition [1–6]. These methods are often considered as feature extraction methods for high-dimensional signals from various application fields, such as transportation, communications, plants, and mines. Unsupervised dimensionality reduction, such as principle component analysis (PCA) [7], does not utilize any label information. Linear discriminant analysis (LDA) is a popular supervised dimensionality reduction method, which derives a projection from simultaneously maximizing the between-class scatter and minimizing the within-class scatter. Semisupervised dimensionality reduction, such as semisupervised discriminant analysis (SDA) [8], makes good use of labeled data while preserving the intrinsic geometric structures of unlabeled data.

In order to handle the data sampled from a low-dimensional manifold, some nonlinear dimensionality reduction methods, such as isometric feature mapping (ISOMAP) [9], locally linear embedding (LLE) [10], and Laplacian Eigenmap (LE) [11], introduce manifold assumption into dimensionality reduction and aim to maximally preserve certain interpoint relationships. But these methods cannot address the out-of-sample extension problem. Thus, locality preserving projections (LPP), as a linear approximation of LE [12], were proposed to both uncover the data manifold and provide out-of-sample extensions. These dimensionality reduction methods could be unified under a framework called graph embedding [13]. To achieve significant improvements, it is feasible to kernelize a certain type of linear methods into nonlinear ones [14–18]. But, the performances of the kernelized versions heavily rely on the selections of kernel functions. With inappropriate kernels, the performances will be degraded and become even worse.

Recently, the advantage of using multiple kernels instead of only one kernel for dimensionality reduction has been demonstrated [15, 19]. Multiple kernel learning for dimensionality reduction (MKL-DR) was proposed to learn an appropriate kernel from the multiple base kernels and a transformation into a lower dimensionality space simultaneously [20]. But, MKL-DR relaxes a nonconvex quadratically constrained quadratic programming (QCQP) into

a semidefinite programming (SDP), which is very time-consuming and has a negative effect on its performance. Recently, a multiple kernel learning method called MKL-TR was proposed to improve the performance of MKL-DR [21]. MKL-TR formulates multiple kernel learning for dimensionality reduction as a trace ratio maximization problem. But both MKL-DR and MKL-TR need to iteratively compute generalized eigendecomposition of dense matrices. Motivated by the efficiency of spectral regression, a fast multiple kernel dimensionality reduction method, termed as MKL-SRTR, was presented to avoid generalized eigendecomposition of dense matrices [22]. It is more efficient than MKL-DR and MKL-TR by virtue of spectral regression. Since MKL-DR, MKL-TR, and MKL-SRTR are all based on graph embedding and manifold assumption, they cannot cope with manifold assumption invalidation. In addition, MKL-DR and MKL-SRTR might be ill-posed if the rank of matrices in their objective functions was not high enough [21].

Since spectral clustering and multiple kernel dimensionality reduction have the same form of optimization based on the manifold assumption, motivated by the spectral embedded clustering framework proposed in [22], we firstly extend the traditional graph embedding framework by incorporating linear regularization terms into its model, termed as extended graph embedding (EGE). Secondly, we introduce multiple kernel learning into EGE (termed as MKL-EGE) to improve the performance of single kernel DR. Compared with traditional multiple kernel dimensionality reduction methods, such as MKL-SRTR, the proposed method not only solves the ill-posed problems but also is more robust against high-dimensional or sparse data. Furthermore, our method directly utilizes a binary search and an alternative optimization scheme to obtain optimal solutions. The experimental results demonstrate that the proposed method achieves better or similar performance compared to other algorithms for supervised, unsupervised, and semisupervised settings.

The remainder of the paper is structured as follows. In Section 2, we briefly introduce the related work. We provide the MKL-EGE framework and the optimization process in Section 3. The experimental results are shown in Section 4. Finally, we give the related conclusions in Section 5. In order to avoid confusion, we give a list of the main notations used in this paper in Notations.

2. Graph Embedding and Its Extension

2.1. Graph Embedding. Specifically, denote an undirected weighted graph by $\mathbf{G} = \{\mathbf{X}, \mathbf{W}\}$, where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T \in \mathbb{R}^{n\times d}$ is a vertex set and $\mathbf{W} \in \mathbb{R}^{n\times n}$ represents an affinity matrix. Each entry W_{ij} of the symmetric matrix $\mathbf{W}$ is the edge weight that characterizes the similarity between a pair of vertices of $\mathbf{G}$. A dimensionality reduction scheme aims at finding a low-dimensional subspace $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n]^T$ ($\mathbf{f}_i \in \mathbb{R}^c$, $c \ll d$) by a complete graph $\mathbf{G}$ whose vertices are over $\mathbf{X}$. The purpose of graph embedding is to represent each vertex of a graph as a low-dimensional vector and preserves similarities between the vertex pairs. The optimal $\mathbf{F}$ could be obtained by solving

$$\begin{aligned} \min_{\mathbf{F}} \quad & \operatorname{tr}\left(\mathbf{F}^T\mathbf{L}\mathbf{F}\right) \\ \text{s.t.} \quad & \operatorname{tr}\left(\mathbf{F}^T\mathbf{D}\mathbf{F}\right) = 1 \text{ or } \operatorname{tr}\left(\mathbf{F}^T\mathbf{L}'\mathbf{F}\right) = 1, \end{aligned} \tag{1}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the graph Laplacian matrix of $\mathbf{G}$ and $\mathbf{D} = \operatorname{diag}(\mathbf{D}_1, \ldots, \mathbf{D}_n)$ is a diagonal matrix with the diagonal elements defined as $\mathbf{D}_i = \sum_{j=1}^{n} \mathbf{w}_{ij}$. $\mathbf{L}' = \mathbf{D}' - \mathbf{W}'$ is the graph Laplacian matrix of another weighted graph $\mathbf{G}'$.

By specifying $\mathbf{W}$ and $\mathbf{D}$ (or $\mathbf{W}$ and $\mathbf{W}'$), the PCA, ISOMAP, LLE, LPP, LDA, local discriminant embedding (LDE), marginal Fisher analysis (MFA) [13], and spectral regression (SR) [23, 24] can all be expressed as graph embedding. Since $\mathbf{L} = \mathbf{D} - \mathbf{W}$ and the constraint $\operatorname{tr}(\mathbf{F}^T\mathbf{D}\mathbf{F}) = 1$ are commonly used, in this paper, we mainly discuss the following form of graph embedding:

$$\begin{aligned} \max_{\mathbf{F}} \quad & \operatorname{tr}\left(\mathbf{F}^T\mathbf{W}\mathbf{F}\right) \\ \text{s.t.} \quad & \operatorname{tr}\left(\mathbf{F}^T\mathbf{D}\mathbf{F}\right) = 1 \end{aligned} \tag{2}$$

which usually relaxes to the following objective function:

$$\max_{\mathbf{F}^T\mathbf{F}=\mathbf{I}_c} \quad \frac{\operatorname{tr}\left(\mathbf{F}^T\mathbf{W}\mathbf{F}\right)}{\operatorname{tr}\left(\mathbf{F}^T\mathbf{D}\mathbf{F}\right)}. \tag{3}$$

2.2. Extended Graph Embedding. The term $\operatorname{tr}(\mathbf{F}^T\mathbf{L}\mathbf{F})$ in problem (4) is actually derived based on the manifold assumption [25]. However, for high-dimensional or sparse data, this assumption may not hold due to the bias caused by the curse of dimensionality. Thus, the low-dimensional manifold structure cannot be exploited by the inaccurate similarity matrix, which would result in the performance degradation of graph embedding.

To address this issue, we try to improve traditional graph embedding framework. Notice that the term $\operatorname{tr}(\mathbf{F}^T\mathbf{L}\mathbf{F})$ can be regarded as the objective function of spectral clustering; we use the spectral embedded clustering method proposed in [26] to extend the graph embedding framework. Specifically, we minimize the following objective function:

$$\min_{\mathbf{F},\mathbf{W},\mathbf{b},\mathbf{F}^T\mathbf{F}=\mathbf{I}_c} \frac{\operatorname{tr}\left(\mathbf{F}^T\mathbf{L}\mathbf{F}\right) + \mu\left(\left\|\mathbf{X}^T\mathbf{W} + \mathbf{1}_n\mathbf{b}^T - \mathbf{F}\right\|^2 + \gamma_g \operatorname{tr}\left(\mathbf{W}^T\mathbf{W}\right)\right)}{\operatorname{tr}\left(\mathbf{F}^T\mathbf{D}\mathbf{F}\right)}, \tag{4}$$

where μ and γ_g are two regularization parameters, $\mathbf{1}_n$ denotes the $n \times 1$ vectors of all 1s, and the second term characterizes the mismatch between the low-dimensional feature matrix $\mathbf{F}$ and the low-dimensional representation of the data.

Theorem 1. *The optimization problem (4) can be transformed into the following minimization problem:*

$$\min_{\mathbf{F}^T\mathbf{F}=\mathbf{I}_c} \frac{\mathrm{tr}\left(\mathbf{F}^T\widetilde{\mathbf{L}}\mathbf{F}\right)}{\mathrm{tr}\left(\mathbf{F}^T\mathbf{D}\mathbf{F}\right)}, \tag{5}$$

where $\widetilde{\mathbf{L}} = \mathbf{L} + \mu\mathbf{L}_s$ *and* $\mathbf{L}_s = \mathbf{I}_n - (1/n)\mathbf{1}_n\mathbf{1}_n^T - \mathbf{X}(\mathbf{X}^T\mathbf{X} + \gamma_g\mathbf{I}_d)^{-1}\mathbf{X}^T$. $\mathbf{I}_n$ *and* $\mathbf{I}_d$ *represent the identity matrix of size n by n and size d by d, respectively.*

Proof. By setting the derivatives of the objective function (4) with respect to $\mathbf{W}$ and $\mathbf{b}$ to zeros, we have

$$\begin{aligned}\mathbf{b} &= \frac{1}{n}\mathbf{F}^T\mathbf{1}_n,\\ \mathbf{W} &= \left(\mathbf{X}^T\mathbf{X} + \gamma_g\mathbf{I}_d\right)^{-1}\mathbf{X}^T\mathbf{F}.\end{aligned} \tag{6}$$

By substituting $\mathbf{W}$ and $\mathbf{b}$ in (4) by (6), the optimization problem (4) becomes

$$\min_{\mathbf{F}^T\mathbf{F}=\mathbf{I}_c} \frac{\mathrm{tr}\left(\mathbf{F}^T\left(\mathbf{L}+\mu\mathbf{L}_s\right)\mathbf{F}\right)}{\mathrm{tr}\left(\mathbf{F}^T\mathbf{D}\mathbf{F}\right)}, \tag{7}$$

where $\mathbf{L}_s = \mathbf{I}_n - (1/n)\mathbf{1}_n\mathbf{1}_n^T - \mathbf{X}(\mathbf{X}^T\mathbf{X} + \gamma_g\mathbf{I}_d)^{-1}\mathbf{X}^T$. This completes the proof of Theorem 1. □

From problem (5), we can find that the form of EGE is similar to that of GE and GE is a special case of EGE when $\mu = 0$. $\widetilde{\mathbf{L}} = \mathbf{L} + \mu\mathbf{L}_s$ can be regarded as a correction of the graph Laplacian matrix $\mathbf{L}$ for high-dimensional data.

Since $\mathbf{L} = \mathbf{D} - \mathbf{W}$, problem (5) can be transformed into the following form:

$$\max_{\mathbf{F}^T\mathbf{F}=\mathbf{I}_c} \frac{\mathrm{tr}\left(\mathbf{F}^T\left(\mathbf{W}-\mu\mathbf{L}_s\right)\mathbf{F}\right)}{\mathrm{tr}\left(\mathbf{F}^T\mathbf{D}\mathbf{F}\right)}. \tag{8}$$

3. Multiple Kernel Learning Based on EGE and Trace Ratio Maximization

Since MKL-DR, MKL-TR, and MKL-SRTR can be viewed as multiple kernel versions of graph embedding, it is natural to establish a multiple kernel learning framework for dimensionality reduction based on EGE.

3.1. Formulation. Suppose the ensemble kernel $\mathbb{K}$ is generated by linearly combining the base kernels $\{\mathbf{K}_m\}_{m=1}^M$; that is, $\mathbb{K} = \sum_{m=1}^M \beta_m\mathbf{K}_m$, where $\beta_m \geq 0$ and $\mathbf{K}_m = \{k_m(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n$. We can find a sample coefficient matrix $\mathbf{A}$ and a kernel weight vector $\boldsymbol{\beta}$ by the following trace ratio optimization problem based on extended graph embedding:

$$\begin{aligned}\max_{\mathbf{A},\boldsymbol{\beta}} \quad & \frac{\mathrm{tr}\left(\mathbf{A}^T\mathbb{K}\left(\mathbf{W}-\mu\mathbf{L}_s\right)\mathbb{K}\mathbf{A}\right)}{\mathrm{tr}\left(\mathbf{A}^T\mathbb{K}\mathbf{D}\mathbb{K}\mathbf{A}\right)}\\ \text{s.t.} \quad & \mathbf{A}^T\mathbf{A} = \mathbf{I}\\ & \beta_m \geq 0, \quad m = 1, 2, \ldots, M,\\ & \sum_m^M \beta_m = 1,\end{aligned} \tag{9}$$

where

$$\begin{aligned}\mathbf{A} &= \left[\boldsymbol{\alpha}_1 \ \ \boldsymbol{\alpha}_2 \ \ \cdots \ \ \boldsymbol{\alpha}_c\right] \in \mathbb{R}^{n\times c},\\ \boldsymbol{\alpha} &= \left[\alpha_1, \ldots, \alpha_n\right]^T \in \mathbb{R}^n,\\ \boldsymbol{\beta} &= \left[\beta_1, \ldots, \beta_M\right]^T \in \mathbb{R}^M,\\ \mathbb{K}^{(i)} &= \begin{bmatrix} k_1(1,i) & \cdots & k_M(1,i)\\ \vdots & \ddots & \vdots\\ k_1(n,i) & \cdots & k_M(n,i)\end{bmatrix} \in \mathbb{R}^{n\times M}.\end{aligned} \tag{10}$$

It should be noted that dimensionality reduction based trace ratio optimization tends to overfitting [27, 28]. To address this issue, a regularization term $\mathrm{tr}(\mathbf{A}^T\lambda\mathbf{I}\mathbf{A})$ is added to the denominator of problem (9) to ensure that $\mathbb{K}\mathbf{D}\mathbb{K} + \lambda\mathbf{I}$ is of full rank. Hence, the objective function could be expressed as follows:

$$\max_{\mathbf{A},\boldsymbol{\beta}} \frac{\mathrm{tr}\left(\mathbf{A}^T\mathbb{K}\left(\mathbf{W}-\mu\mathbf{L}_s\right)\mathbb{K}\mathbf{A}\right)}{\mathrm{tr}\left(\mathbf{A}^T\left(\mathbb{K}\mathbf{D}\mathbb{K} + \lambda\mathbf{I}\right)\mathbf{A}\right)} \tag{11}$$

$$\text{s.t.} \quad \mathbf{A}^T\mathbf{A} = \mathbf{I} \tag{12}$$

$$\beta_m \geq 0, \quad m = 1, 2, \ldots, M, \tag{13}$$

$$\sum_m^M \beta_m = 1. \tag{14}$$

Compared with MKL-SRTR, the proposed method is based on the extended graph embedding framework. Thus, it has more robustness against high-dimensional or sparse data. In addition, our method avoids ill-posed problems.

3.2. Method. To optimize our objective function, the following function that satisfies constraints (13)–(15) is defined:

$$\begin{aligned} f(r) &= \max_{\mathbf{A},\boldsymbol{\beta}} \ s\left(r, \mathbf{A}, \boldsymbol{\beta}\right)\\ &= \max_{\mathbf{A},\boldsymbol{\beta}} \ \mathrm{tr}\left(\mathbf{A}^T\left(\mathbb{K}\left(\mathbf{W} - \mu\mathbf{L}_s\right)\mathbb{K} - r\mathbb{K}\mathbf{D}\mathbb{K} - r\lambda\mathbf{I}\right)\mathbf{A}\right)\\ &= \max_{\mathbf{A},\boldsymbol{\beta}} \ \mathrm{tr}\left(\mathbf{A}^T\left(\mathbb{K}\left(\mathbf{W} - \mu\mathbf{L}_s - r\mathbf{D}\right)\mathbb{K} - r\lambda\mathbf{I}\right)\mathbf{A}\right).\end{aligned} \tag{15}$$

The optimal value of the objective function in (15) is the root of the function $f(r) = \max_{\mathbf{A}^T\mathbf{A}=\mathbf{I}} \mathrm{tr}(\mathbf{A}^T(\mathbb{K}(\mathbf{W} - \mu\mathbf{L}_s - r\mathbf{D})\mathbb{K} - r\lambda\mathbf{I})\mathbf{A})$ [27, 28]. Based on (15), we update r, $\mathbf{A}$, and $\boldsymbol{\beta}$ alternately.

On Optimizing $\mathbf{A}$ *and* r. By fixing $\boldsymbol{\beta}$, optimization problem (11) is simplified to

$$\max_{\mathbf{A}^T\mathbf{A}=\mathbf{I}} \frac{\mathrm{tr}\left(\mathbf{A}^T\mathbf{S}_1\mathbf{A}\right)}{\mathrm{tr}\left(\mathbf{A}^T\mathbf{S}_2\mathbf{A}\right)}, \tag{16}$$

where

$$\mathbf{S}_1 = \mathbb{K}\left(\mathbf{W} - \mu\mathbf{L}_s\right)\mathbb{K}, \tag{17}$$

$$\mathbf{S}_2 = \mathbb{K}\mathbf{D}\mathbb{K} + \lambda\mathbf{I}. \tag{18}$$

Thus, a binary search (giving a lower bound and an upper bound) is used to seek r^* such that $f(r^*) = 0$. The value of $f(r)$ can be easily calculated as the sum of the first c largest eigenvalues of $\mathbf{S}_1 - r\mathbf{S}_2$. Optimal $\mathbf{A}^*$ is finally obtained by performing the eigenvalue decomposition of $\mathbf{S}_1 - r^*\mathbf{S}_2$.

On Optimizing $\boldsymbol{\beta}$. By fixing $\mathbf{A}$ and r, $\boldsymbol{\beta}$ can be obtained by solving the following optimization problem:

$$\begin{aligned} \max_{\boldsymbol{\beta}} \quad & \mathrm{tr}\left(\mathbf{A}^T\mathbb{K}\left(\mathbf{W} - \mu\mathbf{L}_s - r\mathbf{D}\right)\mathbb{K}\mathbf{A}\right), \\ \text{s.t.} \quad & \beta_m \geq 0, \quad m = 1, 2, \ldots, M, \\ & \sum_m^M \beta_m = 1. \end{aligned} \tag{19}$$

We define a function with given $\mathbf{A}$ and r as follows:

$$\mathbf{Q}\left(\boldsymbol{\beta}\right) = \mathrm{tr}\left(\mathbf{A}^T\mathbb{K}\left(\mathbf{W} - \mu\mathbf{L}_s - r\mathbf{D}\right)\mathbb{K}\mathbf{A}\right) \tag{20}$$

and we have

$$\frac{\partial\mathbf{Q}}{\partial\beta_m} = \mathrm{tr}\left(2\mathbf{A}^T\mathbf{K}_m\left(\mathbf{W} - \mu\mathbf{L}_s - r\mathbf{D}\right)\mathbb{K}\mathbf{A}\right). \tag{21}$$

Thus, $\boldsymbol{\beta}$ can be determined by updating the projections of $\boldsymbol{\beta}$ in the direction of $\partial\mathbf{Q}/\partial\boldsymbol{\beta}$. Finally, we define a quadratic programming to satisfy the constraint $\sum_m^M \beta_m = 1$ as

$$\mathbf{G}\left(\boldsymbol{\beta}\right) = \underset{\mathbf{h}^T\mathbf{1}=1, \mathbf{h}\geq\mathbf{0}}{\arg\min} \left\|\mathbf{h} - \boldsymbol{\beta}\right\|_2^2, \tag{22}$$

where $\mathbf{1}$ denotes $n \times 1$ unit vector.

3.3. Algorithms. The proposed algorithm based on EGE and regularized trace ratio, termed as MKL-EGE, is described in Algorithm 1. As can be seen from Algorithm 1, MKL-EGE utilizes a binary search in inner iterations to speed up convergence and adopts updating $\mathbf{A}$ and $\boldsymbol{\beta}$ alternately in outer iterations to seek optimal solutions. Since the proposed algorithm cannot guarantee obtaining the optimal solution r^* exactly, we terminate it within a maximum iteration and choose the best result.

Table 1: Description of experimental datasets.

Datasets	Dimensions	# of samples	# of classes
Ionosphere	33	351	2
Sonar	60	208	2
USPS	256	3000	10
Isolet	617	900	3
MINIST	784	600	3
Yale	1024	165	15
PIE	1024	4080	68
ORL	1024	400	40
COIL-20	1024	1440	20
20NG (comp)	28299	4852	5
20NG (rec)	24990	3968	4
20NG (sci)	30383	3945	4
20NG (talk)	29426	3250	4

3.4. Computational Complexity. For MKL-EGE, the computational complexity of inner iterations is $O(\mathrm{iter}_2 n^3)$, where iter_2 is maximum number of inner iterations. Thus, the computational complexity of the whole algorithm is $O(\mathrm{iter}_1(\mathrm{iter}_2 n^3 + n^3))$, where iter_1 is maximum number of outer iterations. MKL-DR needs to solve the SDP problem in each iteration, which is as high as $O(n^{6.5})$ [20].

The computational complexity of MKL-TR decreases to $O(\mathrm{iter}_1(\mathrm{iter}_2(cn^2 + n^3) + n^3))$ [21]. Since MKL-EGE only needs a small number of iterations to converge, the computational complexity of our method is much lower than that of MKL-DR and MKL-TR.

3.5. Unseen Sample Embedding. After accomplishing the training procedure of MKL-EGE, we can project a new sample $\mathbf{v}$ into the learned subspace by

$$\begin{aligned} &\mathbf{v} \longrightarrow \mathbf{A}^T\mathbb{K}^{(\mathbf{v})}\boldsymbol{\beta}, \\ &\text{where } \mathbb{K}^{(\mathbf{v})} \in \mathbb{R}^{n\times M}, \ \mathbb{K}^{(\mathbf{v})}(i, m) = k_m\left(\mathbf{x}_i, \mathbf{v}\right). \end{aligned} \tag{23}$$

4. Experiments

We compared the proposed MKL-EGE algorithm with MKL-DR [20], MKL-TR [21], and MKL-SRTR [22] on UCI datasets (Sonar, Ionosphere, and Isolet), face recognition datasets (Yale, PIE, and ORL), digits recognition datasets (USPS and MNIST), object recognition datasets (COIL-20), and text datasets (20 newsgroups). We randomly selected 300 samples from each digit for the USPS dataset and used digits 3, 6, and 8 for the MNIST dataset. For 20 newsgroups datasets, four largest topics (comp, rec, sci, and talk) were selected as high-dimensional datasets. For all datasets, we randomly selected samples to form training and testing sets with ratio 1 : 1. The basic information of datasets is listed in Table 1. All the experiments were performed in MATLAB R2013a running in a 3.10 GHZ Intel Core™ i5-2400 with 4-GB RAM.

For all datasets, we first normalized the values of the data vector to the range $[0, 1]$ and used 10 RBF base kernels, whose σ values are set as 0.10, 0.20, 0.40, 0.80,

Input: The matrix of data points $\mathbf{X}, \mathbf{K}_i, \mathbf{D}, \mathbf{W}$, the number of classes c, step length t_1, maximum number of iterations iter_1, parameters λ, μ and γ_g, an error constant ε.
Output: $\mathbf{A}, \boldsymbol{\beta}$
(1) Initialize $\boldsymbol{\beta} = \mathbf{1}/M$, construct the weighted matrix $\mathbf{W}$, calculate $\mathbf{L}_s$.
(2) **Repeat**
(2.1) Calculate $\mathbf{S}_1 = \mathbb{K}(\mathbf{W} - \mu\mathbf{L}_s)\mathbb{K}$ and $\mathbf{S}_2 = \mathbb{K}\mathbf{D}\mathbb{K} + \lambda\mathbf{I}$.
(2.2) Find $\mathbf{p}_1, \ldots, \mathbf{p}_c$ as the first c largest eigenvalues of $\mathbf{S}_1$ and $\mathbf{q}_1, \ldots, \mathbf{q}_c$ as the first c smallest eigenvalues of $\mathbf{S}_2$.
(2.3) Let $r_1 = \text{tr}(\mathbf{S}_1)/\text{tr}(\mathbf{S}_2)$, $r_2 = \sum_{i=1}^{c} \mathbf{p}_i / \sum_{i=1}^{c} \mathbf{q}_i$ and $r = (r_1 + r_2)/2$.
(2.4) **while** $r_1 - r_2 > \varepsilon$ do
(2.4.1) Compute $f(r)$ as the sum of the first c largest eigenvalues of $\mathbf{S}_1 - r\mathbf{S}_2$.
(2.4.2) **If** $f(r) > 0$ **then** $r = r_1$ **else** $r = r_2$.
(2.4.3) $r = (r_1 + r_2)/2$.
(2.5) Obtain $\mathbf{A} = [\boldsymbol{\alpha}_1\ \boldsymbol{\alpha}_2 \cdots \boldsymbol{\alpha}_c]$, where $\boldsymbol{\alpha}_1\ \boldsymbol{\alpha}_2 \cdots \boldsymbol{\alpha}_c$ are the c eigenvectors corresponding to the c largest eigenvalues of $\mathbf{S}_1 - r\mathbf{S}_2$.
(2.6) Set $\boldsymbol{\beta} = (\boldsymbol{\beta} + t_1(\partial\mathbf{Q}/\partial\boldsymbol{\beta}))/\|\partial\mathbf{Q}/\partial\boldsymbol{\beta}\|$.
(2.7) Update $\boldsymbol{\beta} = \mathbf{G}(\boldsymbol{\beta})$.
until iter_1 reached
(3) Output $\mathbf{A}, \boldsymbol{\beta}$, calculate the embedding result $\mathbf{F} = \mathbf{A}^T\mathbb{K}$

Algorithm 1: The proposed MKL-EGE algorithm.

Table 2: Classification accuracy of different DR methods.

Datasets	MKL-DR	MKL-TR	MKL-SRTR	MKL-EGE
Ionosphere	91.07 ± 1.69	**95.17 ± 0.89**	94.50 ± 3.67	94.64 ± 0.97
Sonar	83.90 ± 4.56	86.68 ± 3.47	**88.75 ± 2.83**	87.35 ± 5.46
USPS	93.45 ± 0.62	93.63 ± 0.53	92.73 ± 0.8	**96.43 ± 0.69**
Isolet	94.72 ± 0.25	96.50 ± 0.19	96.80 ± 0.11	**97.82 ± 0.19**
MINIST	91.29 ± 0.99	92.13 ± 0.92	92.64 ± 0.81	**96.13 ± 0.74**
Yale	76.54 ± 5.98	79.83 ± 4.19	78.83 ± 4.31	**82.83 ± 4.72**
PIE	92.41 ± 0.65	94.96 ± 0.11	94.83 ± 0.25	**97.97 ± 0.02**
ORL	90.94 ± 2.20	95.81 ± 1.02	94.63 ± 1.38	**96.32 ± 0.94**
COIL-20	91.87 ± 0.69	93.62 ± 0.35	92.55 ± 0.69	**95.70 ± 0.29**
20NG (comp)	86.00 ± 0.86	84.58 ± 1.05	85.48 ± 0.65	**89.73 ± 0.79**
20NG (rec)	94.02 ± 7.28	96.01 ± 0.66	95.87 ± 2.61	**97.85 ± 0.53**
20NG (sci)	95.19 ± 0.59	95.89 ± 0.60	96.36 ± 1.25	**98.21 ± 0.73**
20NG (talk)	91.22 ± 1.12	92.4 ± 0.98	93.24 ± 2.23	**95.78 ± 0.69**

1.60, 3.20, 6.40, 12.80, 25.60, and 51.20, respectively. In all experiments, we set $t\_1 = 0.5$ and $\varepsilon = 0.001$. The parameter k of k-nearest-neighbor graph is set to 5 empirically. For fair comparisons, we set the parameters λ as 0.5 for MKL-TR and MKL-EGE. For MKL-SRTR and MKL-EGE, we conducted a search of the optimal parameters γ, γ_g, and μ on $\{10^{-9}, 10^{-6}, 10^{-3}, \ldots, 10^{0}, \ldots, 10^{3}, 10^{6}, 10^{9}\}$ by using fivefold cross-validation and reported the best experimental results.

4.1. Experiments on Supervised Learning. The maximum number of iterations for all algorithms is set as 20. For MKL-DR, MKL-SRTR, and MKL-EGE, the affinity matrix $\mathbf{W} = [w_{ij}]$ is defined as

$$w_{ij} = \begin{cases} \dfrac{1}{n_{y_i}}, & \text{if } y_i = y_j, \\ 0, & \text{otherwise.} \end{cases} \tag{24}$$

For MKL-TR, we set $\mathbf{M} = \mathbf{H}(\mathbf{H}^T\mathbf{H})^{\dagger}\mathbf{H}^T - (\mathbf{1}/\mathbf{n})\mathbf{1}\mathbf{1}^T$ and $\mathbf{N} = \mathbf{I} - \mathbf{H}(\mathbf{H}^T\mathbf{H})^{\dagger}\mathbf{H}^T$, where $\mathbf{H}$ represents the indicator matrix with $\mathbf{H}_{ij} = 1$ if $\mathbf{x}_i$ belongs to class j and 0 otherwise. For MKL-DR, the elements of another affinity matrix $\mathbf{W}'$ are all set as $1/\mathbf{N}$. The final reduced dimension is $c - 1$ for all algorithms. We used libSVM [29] with linear kernel to classify the embedding data. All experiments were independently carried out over 20 times.

The mean classification accuracies and the standard deviations of different algorithms are displayed in Table 2. As can be seen from Table 2, MKL-EGE significantly outperforms MKL-DR, MKL-TR, and MKL-SRTR in most datasets, which achieves 11 best recognition rates among all 13 datasets. In particular, the performance of MKL-EGE is much better than that of other algorithms on high-dimensional datasets such as Yale, PIE, ORL, and COIL-20. This is due to the fact that MKL-EGE incorporates EGE and linear regularization terms into its model, which is effective for handling high-dimensional data and can avoid overfitting. Consequently,

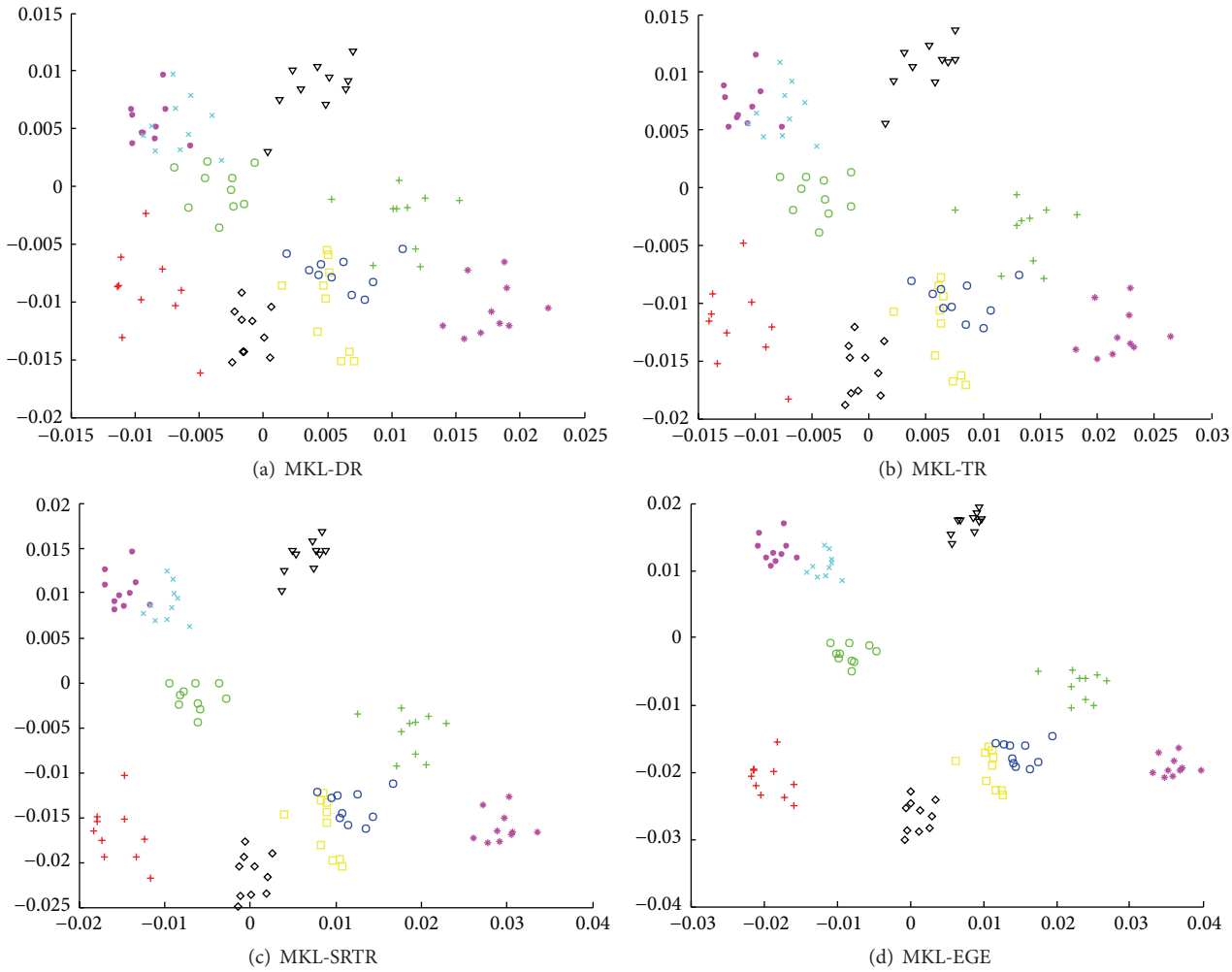


Figure 1: The two-dimensional visualizations of the embedding data from the first 10 classes of ORL. (a) The embedding data in the MKL-DR subspace; (b) the embedding data in the MKL-TR subspace; (c) the embedding data in the MKL-SRTR subspace; (d) the embedding data in the MKL-EGE subspace.

MKL-EGE is more robust than other algorithms based on traditional graph embedding. In addition, the performance of MKL-EGE is very close to that of MKL-TR and MKL-SRTR on low-dimensional dataset, such as Ionosphere, which shows that the proposed method is effective for both low-dimensional and high-dimensional data. The performance of MKL-DR is worst among all algorithms, which validates that the SDP relaxation technique applied in MKL-DR has a negative influence on the performance of dimensionality reduction. The performance of MKL-TR is similar to that of MKL-SRTR, since MKL-SRTR only utilizes spectral regression to improve the speed of MKL-TR.

We used all samples from each class of ORL as training data and used different algorithms to obtain corresponding two-dimensional embedding results. To further validate and compare the final results among different algorithms, we also tested them on PIE, which has the maximum number of samples. The final embedding results are shown in Figures 1 and 2, respectively. As can be seen from Figures 1 and 2, the embedding data obtained by MKL-DR, MKL-TR, and MKL-SRTR is overlapped more seriously than MKL-EGE. The embedding data obtained by MKL-EGE has the best separability, which demonstrates that MKL-EGE is more effective than other algorithms for high-dimensional face data. Consequently, the performance of classification using SVM based on MKL-EGE is best compared to other algorithms.

To compare the computational time of different algorithms, we used all data samples of each dataset as training data to perform different multiple kernel dimensionality reduction methods. The results are displayed in Figure 3. From Figure 3, we can see that MKL-SRTR and MKL-EGE are much faster than MKL-DR and MKL-TR. Since MKL-EGE utilizes a binary search in inner iterations to speed up convergence, its speed is only a little slower than that of MKL-SRTR for the sake of eigenvalue decomposition of dense matrices. The convergence curves of MKL-EGE and MKL-SRTR are displayed in Figure 4. As can be seen from Figure 4, the speed of convergence for MKL-EGE is faster than that of MKL-SRTR; this is due to the fact that

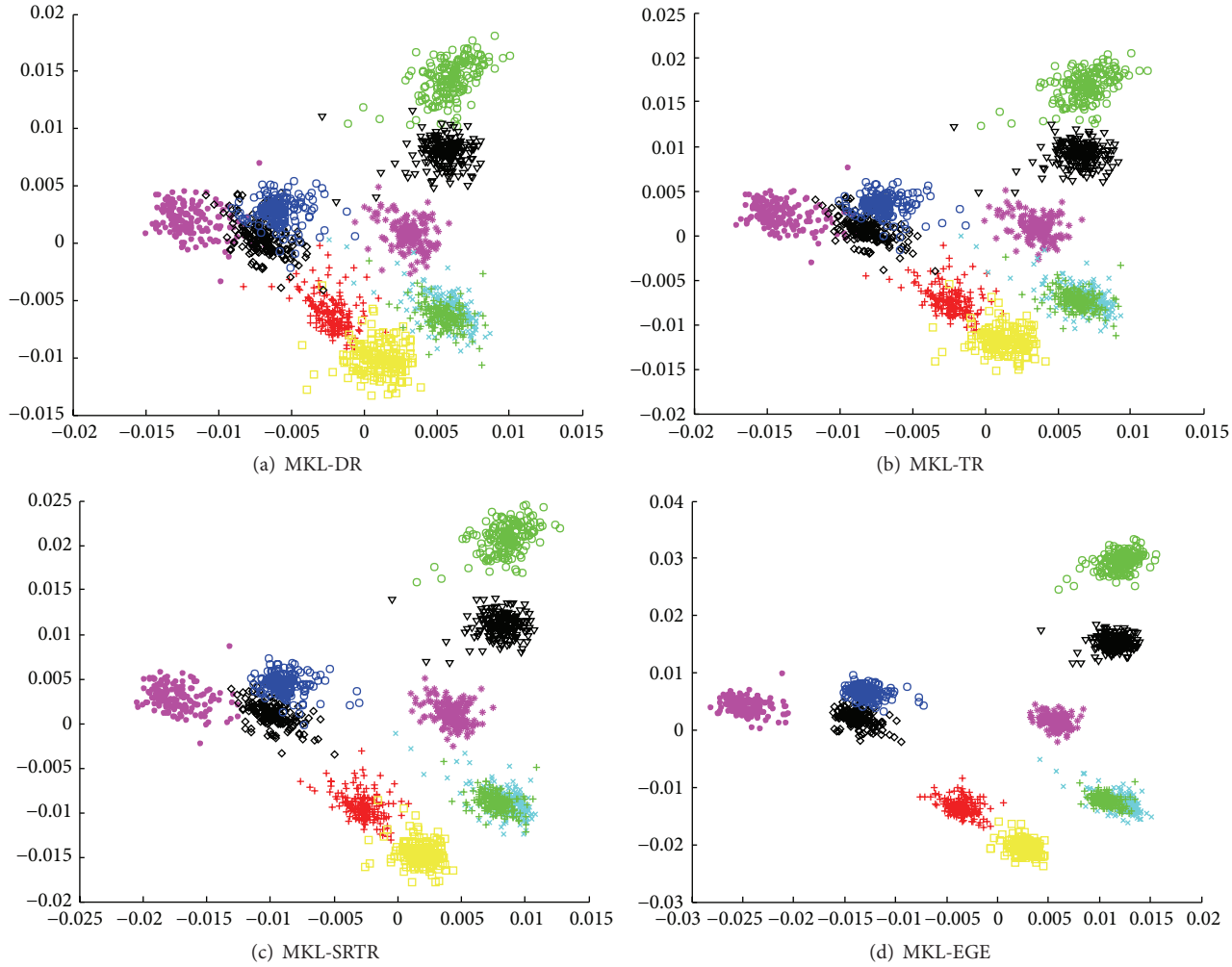


Figure 2: The two-dimensional visualizations of the embedding data from the first 10 classes of PIE. (a) The embedding data in the MKL-DR subspace; (b) the embedding data in the MKL-TR subspace; (c) the embedding data in the MKL-SRTR subspace; (d) the embedding data in the MKL-EGE subspace.

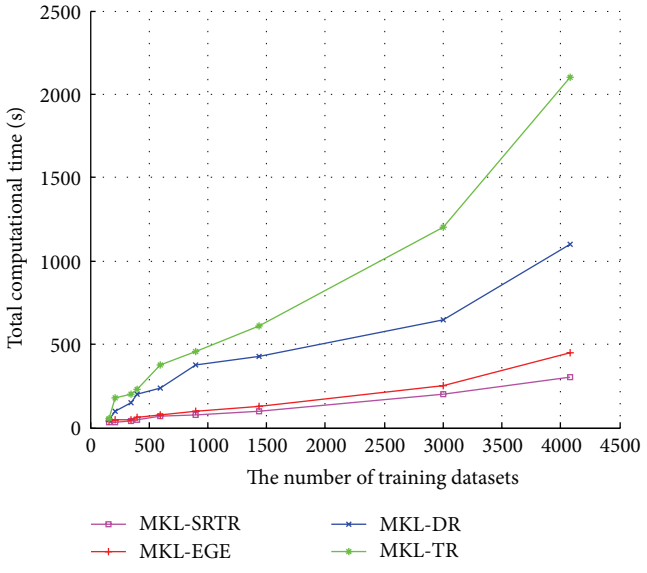


Figure 3: Time cost of different methods on all datasets versus different number of data examples.

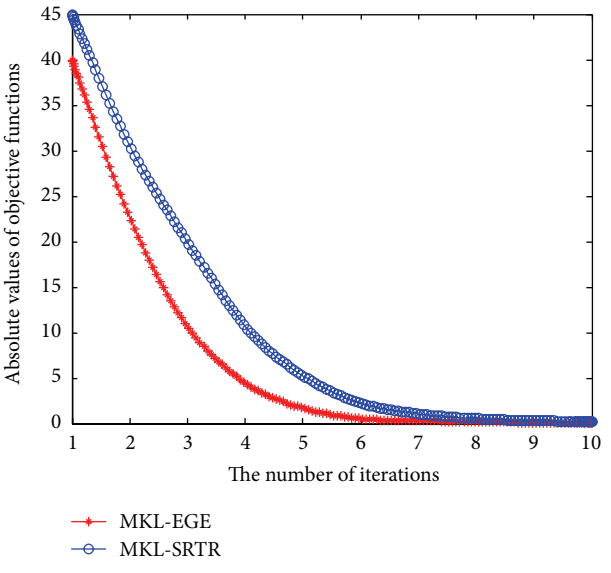


Figure 4: The convergence curves of MKL-EGE and MKL-SRTR.

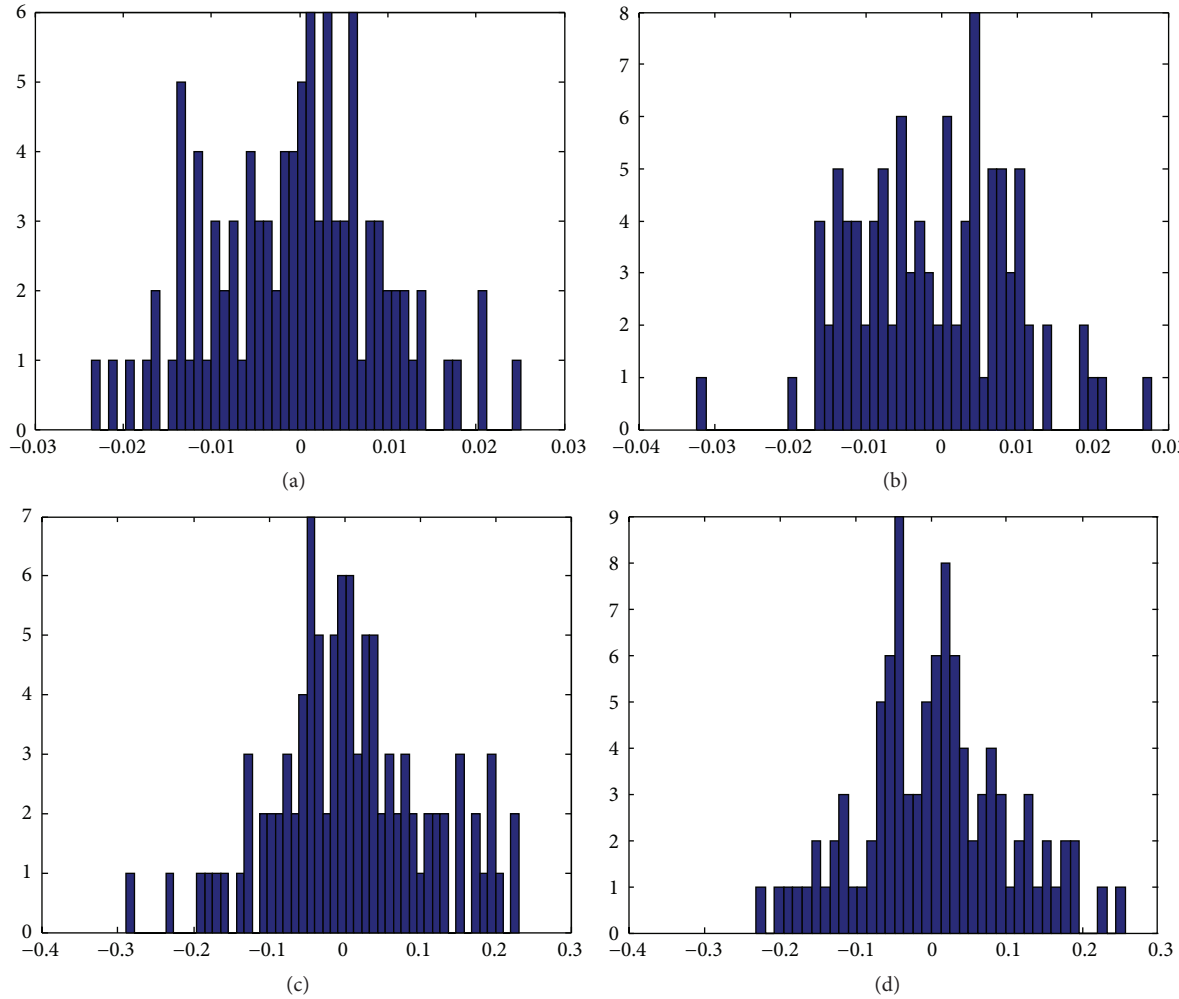


Figure 5: Histograms of the values of $f(r)$ found by different algorithms on USPS. (a) MKL-DR; (b) MKL-TR; (c) MKL-SRTR; (d) MKL-EGE.

MKL-SRTR needs to predefine step length of parameter r and does not adjust adaptively step length in each iteration. For comparing the approximation performances of different algorithms, Figure 5 shows the histograms of the $f(r)$ values obtained by all algorithms in 100 runs. As can be seen from Figure 5, compared with other algorithms, the approximate solutions of MKL-EGE are more concentrated near zero, which validates that our algorithm can more effectively find the root r^* approximately. Overall, the proposed method is the most cost-efficient among all algorithms.

4.2. Experiments on Unsupervised Learning. To evaluate the performance of MKL-EGE in unsupervised settings, we first used all algorithms to project the original data onto a subspace, where the normalized cut spectral clustering (NC) [30] algorithm was performed to evaluate the clustering performance. For MKL-TR, we set $\mathbf{M} = \mathbf{W}$ and $\mathbf{N} = \mathrm{diag}(\mathbf{W1})$, where $\mathbf{W}$ is the affinity matrix for MKL-EGE, MKL-SRTR, and MKL-DR. In the unsupervised case, we set the number of clusters as the number of classes c in each dataset. In order to evaluate the clustering performance, the normalized mutual information (**NMI**) and Rand index (**RI**) [31] were adopted.

We used the same datasets and the same preprocessing procedure as in supervised learning experiments. For unsupervised MKL-DR, initializing $\mathbf{A}$ first obtained more stable performances. Thus, this strategy was adopted in the experiments. To obtain stable results, for each dataset, we computed the average results of each algorithm over 20 runs.

The values of **NMI** and **RI** obtained by these algorithms are reported in Tables 3 and 4, respectively. From Tables 3 and 4, we can see that MKL-EGE performs better than other algorithms in most datasets, which demonstrates that it can improve the performance of dimensionality reduction by using EGE and regularization terms. Consequently, it has the ability to find a more effective combination of base kernels in unsupervised settings. MKL-TR and MKL-SRTR evidently outperform MKL-DR, which indicates that the SDP relaxation used in MKL-DR also has a negative effect on the performance of dimensionality reduction in this case.

Table 3: NMI of different dimensionality reduction algorithms for the clustering task.

Datasets	MKL-TR	MKL-SRTR	MKL-DR	MKL-EGE
Ionosphere	0.1336	0.1422	0.1292	**0.1829**
Sonar	0.0045	0.0004	0.0004	**0.0047**
USPS	0.5725	**0.6069**	0.5451	0.5771
Isolet	0.9474	0.9332	0.9237	**0.9612**
MINIST	0.7287	0.7342	0.7083	**0.7957**
Yale	0.5414	0.5989	0.5212	**0.6305**
PIE	0.0390	0.0629	0.0351	**0.1426**
ORL	0.7813	0.7753	0.7588	**0.8061**
COIL-20	**0.7095**	0.6995	0.6859	0.7034
20NG (comp)	0.3224	0.3173	0.3094	**0.5536**
20NG (rec)	0.7797	0.7497	0.7241	**0.8502**
20NG (sci)	0.7468	0.7372	0.6205	**0.8147**
20NG (talk)	0.3954	0.3489	0.3328	**0.5829**

Table 4: RI of different dimensionality reduction algorithms for the clustering task.

Datasets	MKL-TR	MKL-SRTR	MKL-DR	MKL-EGE
Ionosphere	0.5740	0.5762	0.5853	**0.6034**
Sonar	**0.5023**	0.5000	0.5000	0.5019
USPS	0.8690	**0.8869**	0.8619	0.8720
Isolet	0.9454	0.9497	0.9340	**0.9668**
MINIST	0.8822	0.8871	0.8870	**0.9087**
Yale	0.8942	0.9028	0.9009	**0.9215**
PIE	0.6338	0.6492	0.6335	**0.6953**
ORL	0.9536	0.9505	0.9517	**0.9798**
COIL-20	0.8945	0.8929	0.8945	**0.9075**
20NG (comp)	0.6883	0.6646	0.6648	**0.7897**
20NG (rec)	0.9249	0.9183	0.9172	**0.9662**
20NG (sci)	0.9072	0.9193	0.8157	**0.9316**
20NG (talk)	0.5655	0.6636	0.6210	**0.7239**

4.3. Experiments on Semisupervised Learning. In the semisupervised case, MKL-DR, MKL-TR, MKL-SRTR, and MKL-EGE are actually the multiple kernel extensions of the semisupervised discriminant analysis (SDA) [32–34]. Given l labeled data $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ and u unlabeled data $\{\mathbf{x}_j\}_{j=l+1}^{l+u}$, SDA can be specified by two affinity matrices $\mathbf{W} = [w_{ij}]$ and $\mathbf{W}' = [w'_{ij}]$, defined as follows [34]:

$$w_{ij} = \begin{cases} \dfrac{1}{n_{y_i}} + \delta \cdot s_{ij}, & \text{if } y_i = y_j, \\ \delta \cdot s_{ij}, & \text{otherwise,} \end{cases}$$
$$w'_{ij} = \begin{cases} \dfrac{1}{l}, & \text{if both } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are labeled,} \\ 0, & \text{otherwise,} \end{cases} \tag{25}$$

where

$$s_{ij} = \begin{cases} 1, & \text{if } i \in \mathcal{N}_k(j) \text{ or } j \in \mathcal{N}_k(i) \\ 0, & \text{otherwise} \end{cases} \tag{26}$$

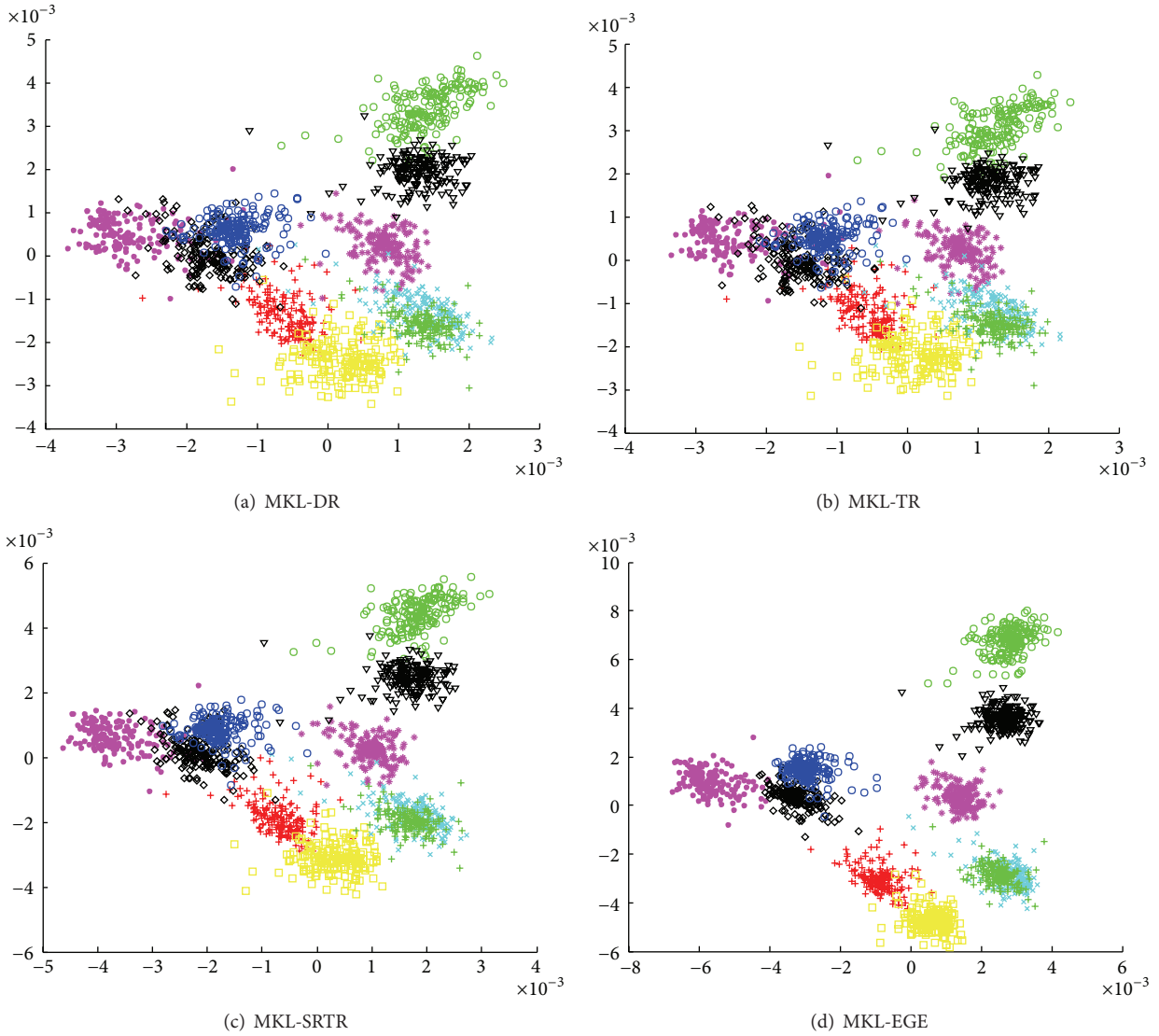
and $0 < \delta \leq 1$ is the parameter to adjust the weight between the label information and unsupervised neighbor information. For MKL-TR, we set $\mathbf{M} = \mathbf{D} - \mathbf{W}$ and $\mathbf{N} = \mathbf{D}' - \mathrm{W}'$. δ is set as 0.1 for all algorithms.

In semisupervised settings, the same datasets and parameter initialization were used. We randomly selected one-half training data as labeled data for each dataset. Each algorithm was independently performed over 20 times. The average classification accuracies as well as the standard deviations are reported in Table 5. As can be seen from Table 5, the proposed MKL-EGE algorithm performs better than MKL-SRTR, MKL-TR, and MKL-DR. Our proposed algorithm, which effectively takes advantage of EGE and regularized trace ratio optimization, can automatically learn weights of base kernels and combine them to improve the performance of dimensionality reduction. By virtue of the same prior information, the proposed algorithm achieves 10 best results among 13 datasets compared with these state-of-the-art methods.

To visualize the semisupervised dimensionality reduction results, we used all samples from the first 10 classes of PIE and projected them into a two-dimensional subspace to generate a graphical representation, shown in Figure 6. From Figure 6, we can observe that the embedding data obtained by MKL-EGE and MKL-SRTR is separated from each other more clearly than MKL-DR and MKL-TR. The embedding data obtained by MKL-EGE has the best separability, which further validates that the performance of MKL-EGE is much better than that of other algorithms in the semisupervised case.

4.4. Experiments on Real World Datasets. To evaluate the effectiveness of MKL-EGE on real world datasets, it serves as a feature extraction method for bearing vibration signals, which were provided by bearing accelerometer sensors under different operating loads and bearing conditions from mines. The vibration signals were collected by using a 16-channel digital audio tape (DAT) recorder at the sampling frequency 12 kHz. Similar to the experimental settings in [35], the experimental vibration data were divided into four datasets, named as D_IRF, D_ORF, D_BF, and D_MIX shown in Table 6, where "07," "14," "21," and "28" mean that fault diameter is 0.007, 0.014, 0.021, and 0.028 inches. We used one-half vibration data as training samples and another one-half as testing samples.

Similar to the experimental settings in [35], we firstly transformed the obtained vibration signals into 10 time domain features, 3 frequency domain features, and 16 time-frequency domain features. Secondly, low-dimensional features were extracted for performing bearings fault diagnosis or prognosis. Finally, SVM was used to evaluate the performance of different DR methods. The first three extracted features corresponding to the largest eigenvalues are employed



(a) MKL-DR (b) MKL-TR

(c) MKL-SRTR (d) MKL-EGE

Figure 6: The two-dimensional visualizations of the embedding data from the first 10 classes of PIE. (a) The embedding data obtained by semisupervised MKL-DR; (b) the embedding data obtained by semisupervised MKL-TR; (c) the embedding data obtained by semisupervised MKL-SRTR; (d) the embedding data obtained by semisupervised MKL-EGE.

Table 5: Classification accuracies of different semisupervised DR methods.

Datasets	MKL-TR	MKL-DR	MKL-SRTR	MKL-EGE
Ionosphere	**82.31 ± 2.82**	80.36 ± 4.33	80.25 ± 3.24	81.02 ± 2.24
Sonar	**61.43 ± 4.45**	55.65 ± 4.87	60.75 ± 4.47	60.83 ± 3.16
USPS	83.58 ± 1.73	81.97 ± 1.25	82.74 ± 0.78	**86.13 ± 0.97**
Isolet	**94.26 ± 0.83**	91.44 ± 0.58	92.54 ± 0.65	93.05 ± 0.71
MNIST	89.68 ± 1.46	88.67 ± 1.43	90.46 ± 1.31	**92.32 ± 1.24**
Yale	35.43 ± 5.32	30.28 ± 4.38	32.29 ± 5.17	**47.08 ± 4.15**
PIE	62.69 ± 6.87	63.31 ± 5.52	64.13 ± 7.84	**68.67 ± 5.65**
ORL	50.13 ± 3.21	46.18 ± 5.19	51.25 ± 2.56	**59.26 ± 3.34**
COIL-20	71.92 ± 2.17	69.47 ± 2.34	70.31 ± 3.29	**75.95 ± 2.56**
20NG (comp)	75.34 ± 0.62	52.92 ± 5.30	76.74 ± 1.25	**82.19 ± 0.76**
20NG (rec)	92.66 ± 0.68	92.23 ± 0.40	93.37 ± 0.73	**94.46 ± 0.89**
20NG (sci)	91.39 ± 0.74	91.88 ± 0.42	91.95 ± 0.26	**93.94 ± 0.54**
20NG (talk)	82.48 ± 0.87	77.63 ± 14.21	84.06 ± 1.59	**88.62 ± 0.69**

Table 6: The experimental datasets.

Datasets	Number	Fault type and diameter	Description
D_IRF	1000	Normal, IRF07, IRF14, IRF21, IRF28	Inner race fault severity
D_ORF	800	Normal, ORF07, ORF14, ORF21	Outer race fault severity
D_BF	1000	Normal, BF07, BF14, BF21, BF28	Ball fault severity
D_MIX	800	Normal, IRF14, ORF14, BF14	Mixed fault classification

Table 7: The classification accuracy rates on four bearing vibration signal datasets.

Datasets	MKL-TR	MKL-DR	MKL-SRTR	MKL-EGE
D_MIX	0.9363	0.9257	0.9485	**0.9835**
D_IRF	0.9415	0.9151	0.9312	**0.9785**
D_ORF	0.9228	0.9238	0.9554	**0.9769**
D_BF	0.9086	0.8992	0.9027	**0.9394**

as the input features of SVM. The classification accuracy rates are reported in Table 7. It can be observed that MKL-EGE achieves much better results compared to other algorithms on all datasets, which further demonstrates the effectiveness of our method for feature extraction of vibration signals in real applications.

5. Conclusion

In this paper, we propose a new multiple kernel dimensionality reduction method called MKL-EGE. By means of EGE and regularized trace ratio maximization, the proposed method not only avoids the SDP relaxation of MKL-DR but improves the performance of multiple kernel dimensionality reduction further. Moreover, the proposed algorithm makes good use of the binary search and alternative optimization scheme to efficiently find optimal solutions. Experimental results validate the effectiveness of this method. In the future, we plan to incorporate pair constraints into our framework and exploit multiple kernel dimensionality reduction via convex optimization.

Notations

R^d: The input d-dimensional Euclidean space
n: The number of total data points
c: The number of classes that the samples belong to
$\mathbf{X}$: $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ is the training data matrix
$\mathbf{Y}$: $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n) \in \mathbb{R}^{m \times n}$ is the 0-1 label vector y_i is the lable of $\mathbf{x}_i$
$k(\mathbf{x}, \mathbf{y})$: Kernel function of data vectors $\mathbf{x}$ and $\mathbf{y}$
$\mathbf{K}$: Kernel matrix $\mathbf{K} = \{k(\mathbf{x}_i, \mathbf{x}_j)\} \in \mathbb{R}^{n \times n}$
$\{\mathbf{K}_m\}_{m=1}^M$: Base kernels
$\boldsymbol{\beta}$: $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_M]^T \in \mathbb{R}^M$, representing nonnegative coefficients of base kernels
$\mathbb{K}$: The ensemble kernel $\mathbb{K} = \sum_{m=1}^M \beta_m \mathbf{K}_m$
$\mathrm{tr}(\mathbf{M})$: The trace of the matrix $\mathbf{M}$, that is, the sum of the diagonal elements of the matrix $\mathbf{M}$.

Competing Interests

The authors declare that there are no competing interests regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (nos. 61403394 and 71573256) and the Fundamental Research Funds for the Central Universities (2014QNA46).

References

[1] L. Li, W. Goh, J. H. Lim, and S. J. Pan, "Extended Spectral Regression for efficient scene recognition," *Pattern Recognition*, vol. 47, no. 9, pp. 2940–2951, 2014.

[2] A. Nazarpour and P. Adibi, "Two-stage multiple kernel learning for supervised dimensionality reduction," *Pattern Recognition*, vol. 48, no. 5, pp. 1854–1862, 2015.

[3] H. Cai, K. Mikolajczyk, and J. Matas, "Learning linear discriminant projections for dimensionality reduction of image descriptors," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 33, no. 2, pp. 338–352, 2011.

[4] X. Huang, Z. Lei, M. Fan, X. Wang, and S. Z. Li, "Regularized discriminative spectral regression method for heterogeneous face matching," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 353–362, 2013.

[5] X. Zhu, Z. Huang, Y. Yang, H. Tao Shen, C. Xu, and J. Luo, "Self-taught dimensionality reduction on the high-dimensional small-sized data," *Pattern Recognition*, vol. 46, no. 1, pp. 215–229, 2013.

[6] X. Zhu, Z. Huang, H. Tao Shen, J. Cheng, and C. Xu, "Dimensionality reduction by Mixed Kernel Canonical Correlation Analysis," *Pattern Recognition*, vol. 45, no. 8, pp. 3003–3016, 2012.

[7] I. T. Jolliffe, *Principal Component Analysis*, Springer, Berlin, Germany, 1986.

[8] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV '07)*, pp. 1–7, Rio de Janeiro, Brazil, October 2007.

[9] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[10] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[11] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in Neural Information Processing Systems*, vol. 14, pp. 585–591, MIT Press, Boston, Mass, USA, 2001.

[12] X. He and P. Niyogi, "Locality preserving projections," in *Proceedings of the Conference in Advances in Neural Information Processing Systems*, pp. 585–591, 2003.

[13] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: a general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40–51, 2007.

[14] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. L. Roux, and M. Ouimet, "Out-of-sample extensions for LLE, isomap, MDS, eigenmaps, and spectral clustering," in *Advances in Neural Information Processing Systems 16*, MIT Press, 2003.

[15] J. Ham, D. D. Lee, S. Mika, and B. Schölkopf, "A kernel view of the dimensionality reduction of manifolds," in *Proceedings of the 21st International Conference on Machine Learning (ICML '04)*, pp. 369–376, Banff, Canada, 2004.

[16] V. De Silva and J. B. Tenenbaum, "Global versus local methods in nonlinear dimensionality reduction," in *Advances in Neural Information Processing Systems 15 (NIPS '02)*, pp. 705–712, MIT Press, 2003.

[17] M. Brand, "Continuous nonlinear dimensionality reduction by kernel eigenmaps," in *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI '03)*, pp. 547–552, 2003.

[18] B. Liu, S.-X. Xia, F.-R. Meng, and Y. Zhou, "Extreme spectral regression for efficient regularized subspace learning," *Neurocomputing*, vol. 149, pp. 171–179, 2015.

[19] C. Cortes, M. Mohri, and A. Rostamizadeh, "Learning nonlinear combinations of kernels," in *Advances in Neural Information Processing Systems*, Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, Eds., vol. 22, pp. 396–404, MIT Press, 2009.

[20] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh, "Multiple kernel learning for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 6, pp. 1147–1160, 2011.

[21] W. Jiang and F.-L. Chung, "A trace ratio maximization approach to multiple kernel-based dimensionality reduction," *Neural Networks*, vol. 49, pp. 96–106, 2014.

[22] M. Liu, W. Sun, and B. Liu, "Multiple kernel dimensionality reduction via spectral regression and trace ratio maximization," *Knowledge-Based Systems*, vol. 83, no. 1, pp. 159–169, 2015.

[23] D. Cai, X. He, and J. Han, "Spectral regression for efficient regularized subspace learning," in *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV '07)*, Rio de Janeiro, Brazil, October 2007.

[24] D. Cai, X. He, W. V. Zhang, and J. Han, "Regularized locality preserving indexing via spectral regression," in *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM '07)*, pp. 741–750, Lisboa, Portugal, November 2007.

[25] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: a geometric framework for learning from labeled and unlabeled examples," *Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.

[26] F. Nie, Z. Zeng, I. W. Tsang, D. Xu, and C. Zhang, "Spectral embedded clustering: a framework for in-sample and out-of-sample spectral clustering," *IEEE Transactions on Neural Networks*, vol. 22, no. 11, pp. 1796–1808, 2011.

[27] T. T. Ngo, M. Bellalij, and Y. Saad, "The trace ratio optimization problem," *SIAM Review*, vol. 54, no. 3, pp. 545–569, 2012.

[28] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang, "Trace ratio vs. ratio trace for dimensionality reduction," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 1–8, Minneapolis, Minn, USA, June 2007.

[29] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, article 27, pp. 1–27, 2011.

[30] S. Xiang, F. Nie, and C. Zhang, "Learning a Mahalanobis distance metric for data clustering and classification," *Pattern Recognition*, vol. 41, no. 12, pp. 3600–3612, 2008.

[31] W. Chen and G. Feng, "Spectral clustering: a semi-supervised approach," *Neurocomputing*, vol. 77, no. 1, pp. 229–242, 2012.

[32] H.-T. Chen, H.-W. Chang, and T.-L. Liu, "Local discriminant embedding and its variants," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 2, pp. 846–853, San Diego, Calif, USA, June 2005.

[33] D. Cai, X. He, and J. Han, "Efficient Kernel Discriminant Analysis via spectral regression," in *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM '07)*, pp. 427–432, Omaha, Neb, USA, October 2007.

[34] D. Cai, X. He, and J. Han, "SRDA: an efficient algorithm for large scale discriminant analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 1, pp. 1–12, 2008.

[35] Z. Xia, S. Xia, L. Wan, and S. Cai, "Spectral regression based fault feature extraction for bearing accelerometer sensor signals," *Sensors*, vol. 12, no. 10, pp. 13694–13719, 2012.