Hindawi
Mathematical Problems in Engineering
Volume 2019, Article ID 6941475, 8 pages
https://doi.org/10.1155/2019/6941475

Research Article

Multiple Kernel Dimensionality Reduction via Ratio-Trace and Marginal Fisher Analysis

Hui Xu,[1] **Yongguo Yang**,[1] **Xin Wang**,[1] **Mingming Liu**,[2] **Hongxia Xie**,[1] **and Chujiao Wang**[1]

[1]*School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, Jiangsu, China*
[2]*School of Intelligent Manufacturing, Jiangsu Vocational Institute of Architectural Technology, Jiangsu, Xuzhou 221008, China*

Correspondence should be addressed to Yongguo Yang; ygyang88@hotmail.com and Mingming Liu; jsjzi_lmm@126.com

Received 12 July 2018; Revised 20 November 2018; Accepted 11 December 2018; Published 14 January 2019

Academic Editor: Ezequiel López-Rubio

Traditional supervised multiple kernel learning (MKL) for dimensionality reduction is generally an extension of kernel discriminant analysis (KDA), which has some restrictive assumptions. In addition, they generally are based on graph embedding framework. A more general multiple kernel-based dimensionality reduction algorithm, called multiple kernel marginal Fisher analysis (MKL-MFA), is presented for supervised nonlinear dimensionality reduction combined with ratio-race optimization problem. MKL-MFA aims at relaxing the restrictive assumption that the data of each class is of a Gaussian distribution and finding an appropriate convex combination of several base kernels. To improve the efficiency of multiple kernel dimensionality reduction, the spectral regression frameworks are incorporated into the optimization model. Furthermore, the optimal weights of predefined base kernels can be obtained by solving a different convex optimization. Experimental results on benchmark datasets demonstrate that MKL-MFA outperforms the state-of-the-art supervised multiple kernel dimensionality reduction methods.

1. Introduction

Recently, multiple kernel dimensionality reduction methods have been attracting many researchers, and a series of methods are proposed based on the graph embedding framework [1–7]. These methods generally transform primal data into high-dimensional feature spaces deduced by a set of base kernels, where a linear transformation is seeking to perform dimensionality reduction. Consequently, these nonlinear dimensionality reduction methods not only deal with high-dimensional data effectively, but automatically select optimal kernels by predefining a set of base kernels. It has been demonstrated that multiple kernel method performs better than single kernel-based methods in dimensionality reduction.

Although existing multiple kernel dimensionality reduction methods are significantly superior to single kernel-based dimensionality reduction methods, they are still confronted with challenging issues. Firstly, these algorithms have to iteratively solve the time-consuming generalized eigenvalue problem, which is a part of alternative optimization methods. Secondly, in addition, it generally transforms the primal model into the simple form by relaxing the SDP (Semidefinite Programming) problem or utilizes gradient descent algorithms to obtain local optima, which all could have a negative effect on its performance. To overcome the shortcomings mentioned above, some multiple kernel dimensionality reduction algorithms based on spectral regression were proposed recently [8–10]. They transform eigen-decomposition of dense matrices into a linear regression problem by means of spectral regression. However, they still make good use of the convex relaxation or gradient descent to optimize the kernel weights. Instead of convex relaxation, a multiple kernel learning framework was recently proposed to avoid relaxing the primal problem [11], which learns a transformation into a space of lower dimension by converting a ratio-trace maximization problem into a semi-infinite linear program. But this method still needs to iteratively compute generalized eigen-decomposition of dense matrices. In addition, these methods are regarded as multiple kernel versions of KDA and

unified under the graph embedding framework. Thus, they all have the assumption that the distribution of each class is considered to be a unimodal Gaussian. This property often does not exist in real world applications and separability of the different classes cannot be well characterized by interclass scatter. Although kernel marginal Fisher analysis (KMFA) has been developed to overcome this limitation by using an intrinsic graph and another penalty graph [12], it has to choose the kernel type and determine its parameters beforehand.

Motivated by these methods, in this paper, a new multiple kernel dimensionality reduction algorithm, called multiple kernel marginal fisher analysis (MKMFA), is presented for supervised nonlinear dimensionality reduction. MKMFA not only solves the problem of the restrictive assumption of existing multiple kernel dimensionality reduction methods, but has the ability of automatically constructing appropriate kernels for nonlinear dimensionality reduction by means of the ratio-trace model. Furthermore, spectral regression is used to address the issue of dense metrics decomposition and speed up the learning of MKMFA. Finally, as other multiple kernel-based dimensionality reduction methods would do, it can also solve the out-of-sample extension problem.

2. Related Work

2.1. Marginal Fisher Analysis. Graph framework is a general platform designing for dimensionality reduction algorithms, and ISOMAP, LLE, and Laplace feature mapping algorithm can be derived from it. With this framework, we develop a new dimensionality reduction algorithm, in order to avoid limitations of traditional linear discriminant analysis in data distribution assumption and available projection direction.

The assumption of the linear discriminant analysis algorithm is that the data of each class is Gaussian distribution, which is usually nonexistent in practical problems. Without this property, the separability of different classes cannot be characterized by interclass scatter. This limitation of LDA can be overcome by developing new standards that are characterized by intraclass compactness and interclass separability. To this end, we propose a new algorithm using the graph embedding framework which is called marginal Fisher analysis (MFA). We design an intrinsic graph with the characteristics of intraclass compactness and another penalty graph characterized by interclass separability. Specifically, the intrinsic graph illustrates the adjacency relationship of the intraclass point, and the connection of each sample to the nearest neighbor k_1 in the same class. The penalty graph describes the adjacency relationship of the interclass marginal point and the marginal point pairs of different categories.

By following the graph embedding formulation, intraclass compactness is characterized from the intrinsic graph by the term [11–13]

$$\begin{aligned} \boldsymbol{S}_c &= \sum_{i\ i\in N_{k_1}^{+}(j)\ or\ j\in N_{k_1}^{+}(i)} \sum \left\| \boldsymbol{w}^T \boldsymbol{x}_i - \boldsymbol{w}^T \boldsymbol{x}_j \right\|^2 W_{ij} \\ &= 2\boldsymbol{w}^T \boldsymbol{X} (\boldsymbol{D} - \boldsymbol{W}) \boldsymbol{X}^T \boldsymbol{w} \end{aligned} \tag{1}$$

where

$$W_{ij} = \begin{cases} 1, & if\ i \in N_{k_1}^{+}(j)\ \ or\ \ j \in N_{k_1}^{+}(i) \\ 0, & else. \end{cases} \tag{2}$$

Here, $N_{k_1}^{+}(i)$ indicates the index set of the k_1 nearest neighbors of the sample $\boldsymbol{x}_i$ in the same class. Interclass separability is characterized by a penalty graph with the term [11–13]

$$\begin{aligned} \boldsymbol{S}_p &= \sum_{i\ (i,j)\in P_{k_2}(c_i)\ or\ (i,j)\in P_{k_2}(c_j)} \sum \left\| \boldsymbol{w}^T \boldsymbol{x}_i - \boldsymbol{w}^T \boldsymbol{x}_j \right\|^2 {W_{ij}}^p \\ &= 2\boldsymbol{w}^T \boldsymbol{X} \left(\boldsymbol{D}^p - \boldsymbol{W}^p \right) \boldsymbol{X}^T \boldsymbol{w} \end{aligned} \tag{3}$$

where

$${W_{ij}}^p = \begin{cases} 1, & if\ \ (i,j) \in P_{k_2}(c_i)\, or\, (i,j) \in P_{k_2}\left(c_j\right) \\ 0, & else. \end{cases} \tag{4}$$

Here, $P_{k_2}(c)$ is a set of data pairs that are the k_2 nearest pairs among the set $\{(i,j), i \in \pi_c, j \notin \pi_c\}$, where π_c denotes the index set of the samples belonging to the cth class. The algorithmic procedure of marginal Fisher analysis algorithm is formally stated as follows [11–13]:

(1) Firstly, project the data set into PCA subspace by preserving $N - N_C$ dimensions or a certain energy. The transformation matrix of PCA was represented by $\boldsymbol{W}_{\mathrm{PCA}}$.

(2) Construct the intraclass compactness and interclass separability graphs. In the intraclass compactness graph, for each sample $\mathrm{x_i}$, set the adjacency matrix $W_{ij} = W_{ji} = 1$ if $\boldsymbol{x}_i$ is among the k_1-nearest neighbors of $\boldsymbol{x}_j$ in the same class. In the interclass separability graph, for each class c, set the similarity matrix $W_{ij}^p = 1$ if the pair (i, j) is among the k_2 shortest pairs among the set$\{(i,j), i \in \pi_c, j \notin \pi_c\}$.

(3) Marginal Fisher Criterion. From the linearization of the graph embedding framework, we have the Marginal Fisher Criterion

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \frac{\boldsymbol{w}^T \boldsymbol{X} (\boldsymbol{D} - \boldsymbol{W}) \boldsymbol{X}^T \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{X} \left(\boldsymbol{D}^p - \boldsymbol{W}^p \right) \boldsymbol{X}^T \boldsymbol{w}} \tag{5}$$

which is a special linearization of the graph embedding framework with

$$\boldsymbol{B} = \boldsymbol{D}^p - \boldsymbol{W}^p \tag{6}$$

(4) Output the final linear projection direction as

$$\boldsymbol{w} = \boldsymbol{W}_{PCA} \boldsymbol{w}^* \tag{7}$$

2.2. Ratio-Trace Optimization Problem. For any two $d \times d$ symmetric positive semidefinite matrices $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$, the ratio-trace problem is defined as [14]

$$\max_{\mathbf{W}} \operatorname{trace} \left[\left(\mathbf{W}^{\mathbf{T}} \mathbf{S}_1 \mathbf{W} \right)^{-1} \left(\mathbf{W}^{\mathbf{T}} \mathbf{S}_2 \mathbf{W} \right) \right] \tag{8}$$

For a given kernel function $\mathbf{K}$, the kernelized versions of these algorithms solve the following ratio-trace problem:

$$\max_{\boldsymbol{A}} \operatorname{trace}\left[\left(\mathbf{A}^{\mathrm{T}}\left((1-\rho)\,\mathbf{KLK}+\rho\mathbf{K}\right)\mathbf{A}\right)^{-1} \cdot\left(\mathbf{A}^{\mathbf{T}}\mathbf{KL}'\mathbf{KA}\right)\right] \tag{9}$$

where $\mathbf{A}$ is a transformation matrix, $\mathbf{K}$ is the $N \times N$ kernel matrix with $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, and $\rho \in (0,1)$ is a regularization parameter used to prevent overfitting. $\mathbf{L}$ and $\mathbf{L}'$ are (algorithm-dependent) $N \times N$ symmetric positive semidefinite matrices. The optimal solution to (9) is given by the generalized eigenvectors corresponding to the nonzero generalized Eigenvalues:

$$\mathbf{KL}'\mathbf{K}\gamma = \lambda\left((1-\rho)\,\mathbf{KLK}+\rho\mathbf{K}\right)\gamma \tag{10}$$

Once $\mathbf{A}$ is obtained, the new representation for a data sample $\boldsymbol{x}_{\text{new}} \in \mathbb{R}^d$ can be computed using

$$\boldsymbol{x}_{\text{new}} = \mathbf{A}^{\mathrm{T}}\left[k\left(\boldsymbol{x}_1, \boldsymbol{x}\right), \ldots, k\left(\boldsymbol{x}_N, \boldsymbol{x}\right)\right]^T \tag{11}$$

3. Multiple Marginal Fisher Analysis Kernel Dimensionality Reduction via Ratio-Trace

3.1. Kernel Marginal Fisher Analysis via Ratio-Trace. The kernel trick is widely used to improve the separation ability of a linear supervised dimensionality reduction algorithm. By using the kernel trick, the marginal Fisher analysis can be further improved. By replacing $\mathbf{L}$ and $\mathbf{L}'$ by $(\boldsymbol{D}^p - \boldsymbol{W}^p)$ and $(\boldsymbol{D}-\boldsymbol{W})$, respectively, problem (9) can be rewritten as follows:

$$\max_{\boldsymbol{A}} \operatorname{trace}\left[\left(\mathbf{A}^{\mathrm{T}}\left((1-\rho)\,\mathbf{K}\left(\boldsymbol{D}-\boldsymbol{W}\right)\mathbf{K}+\rho\mathbf{K}\right)\mathbf{A}\right)^{-1} \cdot\left(\mathbf{A}^{\mathbf{T}}\mathbf{K}\left(\boldsymbol{D}^p-\boldsymbol{W}^p\right)\mathbf{KA}\right)\right] \tag{12}$$

Note that the graphs of kernel marginal Fisher analysis (KMFA) may be different from MFA because the nearest neighbors k_1 for each sample in KMFA is different from one in MFA. In each class the k_1 nearest in-class neighbors of each sample and the k_2 closest out-of-class sample pairs can be measured through the use of the kernel mapping function $\Phi(\boldsymbol{x})$ from the original feature space to the higher dimensional Hilbert space. The distance between sample $\boldsymbol{x}_i$ and sample $\boldsymbol{x}_j$ can be obtained by the following formula:

$$\begin{aligned} D\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) &= \left\|\Phi\left(\boldsymbol{x}_i\right) - \Phi\left(\boldsymbol{x}_j\right)\right\| \\ &= \sqrt{k\left(\boldsymbol{x}_i, \boldsymbol{x}_i\right) + k\left(\boldsymbol{x}_j, \boldsymbol{x}_j\right) - 2k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)} \end{aligned} \tag{13}$$

3.2. Multiple Kernel Marginal Fisher Analysis Dimensionality Reduction. In this section, a multiple kernel Fisher analysis framework is presented to incorporate spectral regression and ratio-trace into multiple kernel learning for dimensionality reduction. On one hand, spectral regression does not increase speed at the cost of some accuracy. On the other hand, the ratio-trace optimization algorithm can avoid conventional convex relaxation or gradient descent optimization method. The formulation of multiple kernel learning with MFA and ratio-trace will be illustrated, which not only combines multiple kernel dimensionality reduction with MFA, but selects optimal kernels more effectively than other multiple kernel dimensionality reduction methods by semi-infinite linear program (SLIP).

In the MKL framework, the kernel function $\mathbb{K}$ is parametrized as a linear combination of predefined base kernels $\boldsymbol{K}_1, \ldots, \boldsymbol{K}_m$:

$$\mathbb{K} = \sum_{m=1}^{M} \beta_m \boldsymbol{K}_m, \quad \beta_m \geq 0, \ \sum_{\mathrm{m}=1}^{\mathrm{M}} \beta_m = 1, \tag{14}$$

where $\boldsymbol{K}_m = \{k_m(\boldsymbol{x}_i, \boldsymbol{x}_j)\}_{i,j=1}^n$ and the weights $\beta = [\beta_1, \ldots, \beta_m]$ are learned from the data. Under the kernel marginal Fisher analysis framework based on ratio-trace, a multiple kernel variant of KMFA is deduced by combining MFA with MKL, termed as MKMFA, which is formulated as the following optimization problem:

$$\underset{\mathrm{A},\mathbb{K},\beta}{\text{maximize}} \operatorname{trace}\left[\left(\mathbf{A}^{\mathbf{T}}\left((1-\rho)\,\mathbb{K}\left(\boldsymbol{D}-\boldsymbol{W}\right)\mathbb{K}+\rho\mathbb{K}\right) \cdot\mathbf{A}\right)^{-1}\left(\mathbf{A}^{\mathbf{T}}\mathbb{K}\left(\boldsymbol{D}^p-\boldsymbol{W}^p\right)\mathbb{K}\mathbf{A}\right)\right] \tag{15}$$

Given the input data point $(\boldsymbol{x}_i, y_i)$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ and y_i is the class label of $\boldsymbol{x}_i$. Denote $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots., \boldsymbol{x}_n] \in \mathbb{R}^{d\times n}$ as the training data matrix. The detailed steps of MKMFA are given as follows:

Step 1. Constructing the intraclass compactness graph $\boldsymbol{W}$ and interclass separability graph $\boldsymbol{W}^p$.

Step 2. We extend the Marginal Fisher Criterion to the multiple kernel case in the following way:

Firstly, intraclass compactness is characterized from the intrinsic graph by the term

$$\begin{aligned} S_c &= \sum_i \sum_{i\in N_{k_1}^+(j) \text{ or } j\in N_{k_1}^+(i)} \left\|\boldsymbol{\alpha}^T\mathbb{K}^{(i)}\boldsymbol{\beta} - \boldsymbol{\alpha}^T\mathbb{K}^{(j)}\boldsymbol{\beta}\right\|^2 \\ &= 2\boldsymbol{\alpha}^T\mathbb{K}\left(\boldsymbol{D}-\boldsymbol{W}\right)\mathbb{K}\boldsymbol{\alpha}, \end{aligned} \tag{16}$$

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]^T \in \mathbb{R}^n$, $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_M]^T \in \mathbb{R}^M$, and $\mathbb{K}^{(i)} = \begin{bmatrix} k_1(1,i) & \cdots & k_M(1,i) \\ \vdots & \ddots & \vdots \\ k_1(n,i) & \cdots & k_M(n,i) \end{bmatrix} \in \mathbb{R}^{n\times M}$. $\boldsymbol{D} = \operatorname{diag}(\boldsymbol{D}_1, \ldots, \boldsymbol{D}_n)$ is a diagonal matrix with the diagonal elements defined as $\boldsymbol{D}_i = \sum_{j=1}^n \boldsymbol{w}_{ij}$.

Secondly, interclass separability is characterized by a penalty graph with the term

$$\begin{aligned} S_p &= \sum_i \sum_{(i,j)\in P_{k_2}(c_i) \text{ or } (i,j)\in P_{k_2}(c_j)} \left\|\boldsymbol{\alpha}^T\mathbb{K}^{(i)}\boldsymbol{\beta} - \boldsymbol{\alpha}^T\mathbb{K}^{(j)}\boldsymbol{\beta}\right\|^2 \\ &= 2\boldsymbol{\alpha}^T\mathbb{K}\left(\boldsymbol{D}^p-\boldsymbol{W}^p\right)\mathbb{K}\boldsymbol{\alpha}, \end{aligned} \tag{17}$$

where $\boldsymbol{D}^p$ is the degree matrix of $\boldsymbol{W}^p$.

To obtain a multidimensional projection, we consider a set of c sample coefficient vectors, denoted by $\boldsymbol{A} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots \boldsymbol{\alpha}_c] \in \mathbb{R}^{n\times c}$. Finally, Multiple Kernel Marginal Fisher Criterion can be denoted as follows:

$$\begin{aligned} \max_{\boldsymbol{A},\boldsymbol{\beta}} \quad & \operatorname{tr}\left[\left(\boldsymbol{A}^{\boldsymbol{T}}\left(\left(1-\rho\right)\mathbb{K}\left(\boldsymbol{D}-\boldsymbol{W}\right)\mathbb{K}+\rho\mathbb{K}\right)\boldsymbol{A}\right)^{-1}\left(\boldsymbol{A}^{\boldsymbol{T}}\mathbb{K}\left(\boldsymbol{D}^{p}-\boldsymbol{W}^{p}\right)\mathbb{K}\boldsymbol{A}\right)\right] \\ \text{s.t.} \quad & \beta_m \geq 0,\ \sum_{m}^{M}\beta_m = 1,\ m = 1,2,\ldots,M \end{aligned} \tag{18}$$

where $\rho \in (0,1)$ is a regularization parameter used to prevent overfitting.

Step 3. Assume the ranks of $(\boldsymbol{D}-\boldsymbol{W})$ and $(\boldsymbol{D}^p-\boldsymbol{W}^p)$ are r and r', respectively. Let $\{(\alpha_i, \boldsymbol{\mu}_i)\}_{i=1}^{r}$ and $\{(\beta_i, \boldsymbol{\nu}_i)\}_{i=1}^{r'}$ be the nonzero Eigenvalue-Eigenvector pairs of $(\boldsymbol{D}-\boldsymbol{W})$ and $(\boldsymbol{D}^p-\boldsymbol{W}^p)$, respectively. We can obtain the optimal $\boldsymbol{\beta}^*$ by solving the following semi-infinite linear program [15]:

$$\begin{aligned} \max_{\tau,\boldsymbol{\beta}} \quad & \tau \\ \text{s.t.} \quad & \sum_{m}^{M}\beta_m S_m\left(\boldsymbol{\theta}\right) \geq \tau, \quad \forall\boldsymbol{\theta} \in \mathbb{R}^{r\times r'}, \\ & \beta_m \geq 0,\ \sum_{m}^{M}\beta_m = 1,\ \forall m \end{aligned} \tag{19}$$

where $S_m(\boldsymbol{\theta}) = (1/\rho)\sum_{i=1}^{r'}(\boldsymbol{\theta}_i{}^T\boldsymbol{\theta}_i/4(1-\rho)+\boldsymbol{\theta}_i{}^T\boldsymbol{P}^T\boldsymbol{K}_m\boldsymbol{P}\boldsymbol{\theta}_i/4\rho-\boldsymbol{\theta}_i{}^T\boldsymbol{P}^T\boldsymbol{K}_m\boldsymbol{q}_i)+\operatorname{trace}(\boldsymbol{K}_m(\boldsymbol{D}^p-\boldsymbol{W}^p))$ being M functions defined $\forall\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{r'}] \in \mathbb{R}^{r\times r'}$ and $\boldsymbol{P} = [\sqrt{\alpha_1}\boldsymbol{\mu}_1, \ldots, \sqrt{\alpha_r}\boldsymbol{\mu}_r]$ and $\boldsymbol{q}_i = \sqrt{\beta_i}\boldsymbol{\nu}_i$ for $i = 1, 2, \ldots, r'$.

Step 4. Solve the ratio-trace problem (18) using spectral regression to get optimal $\boldsymbol{A}^*$. Since $(\boldsymbol{D}^p-\boldsymbol{W}^p)$ and $(\boldsymbol{D}-\boldsymbol{W})$ are all sparse matrices, we can use spectral regression to obtain $\boldsymbol{A}$ in the following way:

(1) Find the largest c generalized eigenvectors $\boldsymbol{y}_0, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{c-1}$ of the following eigen-problem:

$$\left(\boldsymbol{D}^p-\boldsymbol{W}^p\right)\boldsymbol{y} = \lambda\left[\left(1-\rho\right)\left(\boldsymbol{D}-\boldsymbol{W}\right)+\rho\boldsymbol{I}\right]\boldsymbol{y}, \tag{20}$$

(2) Find $\boldsymbol{\alpha}_k (k = 1 \ldots c)$ by solving the following least squares regression:

$$\boldsymbol{\alpha}_k = \arg\min_{\boldsymbol{\alpha}}\left(\sum_{i=1}^{n}\left(\boldsymbol{\alpha}^T\mathbb{K}^{(i)}\boldsymbol{\beta}^* - \boldsymbol{y}_i^k\right)^2 + \gamma\left\|\boldsymbol{\alpha}\right\|^2\right) \tag{21}$$

where $\boldsymbol{y}_i^k$ is the i-th element of $\boldsymbol{y}^k$.

Algorithm 1 summarizes the algorithm for solving (18). This iterative algorithm is referred to as MKL-MFA. The alternating algorithm for solving the proposed SILP problem belongs to a family of algorithms for solving general semi-infinite programming problems called the exchange methods, in which the constraints are exchanged at each iteration. These methods have been guaranteed to converge [16].

Compared with existing supervised multiple kernel dimensionality reduction based on LDA [14, 17–23], the proposed method has the following advantages:

(1) The projection direction that MFA can use is much greater than that of LDA, and the dimension size is determined k_2, that is, the number of the shortest pairs between in-class and out-of-class samples.

(2) Do not assume the data distribution in each class, and the intraclass compactness is characterized by the sum of the distance between each data and the nearest neighbor k_1 in the same class. Therefore, discriminant analysis is more general.

(3) Without prior information of data distribution, the interclass margin can better characterize the separability in different classes than the interclass variance in LDA.

(4) Avoid conventional convex relaxation or gradient descent optimization algorithm, and optimal kernels can be more effectively obtained than other multiple kernel dimensionality reduction methods.

4. Experiments

To validate the effectiveness of the proposed method, all algorithms are carried out on UCI (University of California, Irvine) datasets, digits recognition and face recognition datasets. The characteristics of the datasets are summarized in Table 1. For fair comparison, the final reduced dimension is equal to the number of classes of each dataset for all algorithms and the libSVM tool is used to classify the reduced data. For each dataset, training and testing sets are selected randomly with ratio 1:1. After that, the values of samples are normalized to the range $[0, 1]$ for each dataset. Finally, we analyze and compare our method with other algorithms by repeating each algorithm 20 runs.

For fair comparison, 10 RBF base kernels are predefined to construct the ensemble kernel $\mathbb{K}$ as the MKL-TR algorithm [12], and the values of parameter σ are set as 0.10, 0.22, 0.46, 1.00, 2.15, 4.46, 10.00, 21.54, 46.42, and 100.00, respectively. Based on these base kernels, MKL-MFA is mainly compared with EMFA[10], MKL-DR [11], MKL-TR [12], and MKL-SRTR [13] in supervised settings. The parameters k_1 and k_2 of MKL-MFA and EMFA are all equal to 10, while the parameters γ and ρ are specified by cross-validation. For MKL-TR, we set $\boldsymbol{M} = \boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{H})^{\dagger}\boldsymbol{H}^T - (\boldsymbol{1}/\boldsymbol{n})\boldsymbol{1}\boldsymbol{1}^{\boldsymbol{T}}$ and $\boldsymbol{N} =$

Input:$\{\boldsymbol{K}_m\}_{\mathrm{m}=1}^{\mathrm{M}}$, T, ρ, ϵ.
Output: $\boldsymbol{A}^*$
Step1: Initialization: $\beta_m = 1/M, \forall m, \zeta = +\infty, t = 1, \mathrm{C} = \boldsymbol{\Phi}$, $(1-\rho)\mathbb{K}(\boldsymbol{D}-\boldsymbol{W})\mathbb{K} + \rho\mathbb{K}, \mathbb{K}(\boldsymbol{D}^p - \boldsymbol{W}^p)\mathbb{K}$.
Step2: Compute $\boldsymbol{P} = [\sqrt{\alpha_1}\boldsymbol{\mu}_1, \ldots, \sqrt{\alpha_r}\boldsymbol{\mu}_r]$, and $\boldsymbol{q}_i = \sqrt{\beta_i}\boldsymbol{\nu}_i$, where $\{(\alpha_i, \boldsymbol{\mu}_i)\}_{i=1}^{r}$ and $\{(\beta_i, \boldsymbol{\nu}_i)\}_{i=1}^{r'}$ are the non-zero Eigenvalue-Eigenvector Pairs of $\mathbb{K}(\boldsymbol{D}^p - \boldsymbol{W}^p)\mathbb{K}$ and $(1-\rho)\mathbb{K}(\boldsymbol{D}-\boldsymbol{W})\mathbb{K} + \rho\mathbb{K}$ respectively.
Step3: Solve the SILP (19) to obtain $\boldsymbol{\beta}^*$ as follows:
While $t \leq \mathrm{T}$
$\mathbb{K} = \sum_{m=1}^{M} \beta_m \boldsymbol{K}_m$
For i = $1, \ldots, \mathrm{r}'$
Compute $\boldsymbol{\eta}_i^*$ by solving $(\boldsymbol{I}/2(1-\rho) + \boldsymbol{P}^T\boldsymbol{K}\boldsymbol{P}/2\rho)\boldsymbol{\eta}_i^* = \boldsymbol{P}^T\boldsymbol{K}\boldsymbol{q}_i$.
end
If $|1 - (\sum_{m=1}^{M} \beta_m S_m(\boldsymbol{\eta}^*))/\zeta| < \epsilon$ break;
else
Add $\boldsymbol{\eta}^*$ to the constraint set C. Update μ and ζ by solving restricted Version of (19) using only $\boldsymbol{\eta} \in \mathrm{C}$.
end
$t = t+1$;
end
Step4: Solve the ratio-trace problem(3) using spectral regression with $\mathbb{K}^* = \sum_{m=1}^{M} \beta_{\mathrm{m}}^* \boldsymbol{K}_m$,
To get optimal $\boldsymbol{A}^*$.
Step5: The new non-linearly transformed representation for a data sample $\mathrm{x} \in \mathbf{R}^{\mathrm{d}}$ can be computed by
$\boldsymbol{x}' = \mathbf{A}^{*\mathrm{T}}[\mathbb{K}^*(\boldsymbol{x}_1, \boldsymbol{x}), \ldots, \mathbb{K}^*(\boldsymbol{x}_N, \boldsymbol{x})]^{\mathrm{T}}$ and $\mathbb{K}^*(\boldsymbol{x}_i, \boldsymbol{x}) = \sum_{m=1}^{M} \beta_m^* \boldsymbol{K}_m(\boldsymbol{x}_i, \boldsymbol{x})$.

Algorithm 1: MKL-MFA algorithm.

Table 1: Description of benchmark 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	680	4
ORL	1024	400	40
COIL-20	1024	1440	20

$\boldsymbol{I} - \boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{H})^{\dagger}\boldsymbol{H}^T$, where $(\bullet)^{\dagger}$ denotes pseudoinverse and $\boldsymbol{H}$ is the indicator matrix with $\boldsymbol{H}_{ij} = 1$ if $\boldsymbol{x}_i$ belongs to class j, and 0 otherwise. As the settings of MKL-SRTR and MKL-DR, we also define the affinity matrix $\boldsymbol{W} = [w_{ij}]$ as

$$w_{ij} = \begin{cases} \dfrac{1}{n_{y_i}}, & if \ \ y_i = y_j, \\ 0, & \text{otherwise}, \end{cases} \tag{22}$$

and set another affinity matrix $\boldsymbol{W}' = [w'_{ij}]$, where $w'_{ij} = 1/\boldsymbol{N}$. For EMFA, the parameter γ and the number of hidden nodes L are obtained by a 10-fold cross-validation. According to the mean classification accuracies and the standard deviations, the performance of each algorithm is evaluated and reported in Table 2.

From Table 2, the performance of MKL-MFA is significantly superior to that of KMFA, EMFA, MKL-TR, MKL-DR, and MKL-SRTR on all datasets. The results of KMFA are worse than other methods; this is due to the fact that the combination of multiple base kernels can improve the performance of single kernel-based algorithms. Our method outperforms MKL-DR, MKL-TR, and MKL-SRTR, since MKL-MFA utilizes the penalty graph to characterize the interclass marginal point adjacency relationship. Without prior information on data distributions, the interclass margin can better characterize the separability of different classes than the interclass variance in KDA. The performance of our model also goes beyond that of EMFA, which can be attributed to the fact that our model finds the global optimum by converting a ratio-trace maximization problem into a semi-infinite linear program instead of the convex relaxation or gradient descent. Experimental results demonstrate that MKL-MFA makes good use of ratio-trace optimization, MKL and MFA to achieve the outstanding discriminant analysis power and yields the best results.

To analyze the performance of our model further, we test MKL-TR and MKL-MFA in the 30 runs of experiments on Ionosphere with different splits of training and testing set. The mean values of kernel weights are reported in Table 3. For comparison, in Table 3 the best classification accuracies of KFDA corresponding to the 10 base kernels are also displayed, respectively. As can be seen from Table 3, K_3, K_4, K_5, and K_6 are more suitable for KFDA than other kernels. It can be observed that our method tends to assign larger weights on K_3, K_4, K_5, and K_6. As a result, our method is able to combine base kernels with appropriate weights and overly outperforms KFDA with the best single kernel.

To validate the effectiveness of our model on high-dimensional data, we select 40 samples from 9 classes of PIE face dataset, projecting them into two-dimensional space

Table 2: Classification accuracy of different DR methods.

Datasets	MKL-DR	MKL-TR	MKL-SRTR	KMFA	EMFA	MKL-MFA
Ionosphere	91.24 ± 3.67	92.43± 0.86	92.62± 0.73	90.84±1.48	93.73±1.42	**95.66±1.52**
Sonar	80.44± 5.46	83.26± 4.13	84.35± 2.58	82.65±2.26	87.52±2.25	**90.75±2.39**
USPS	90.65± 0.77	92.93± 0.43	92.82± 0.52	91.69±0.93	96.59±0.37	**97.83±0.45**
Isolet	93.47± 0.22	95.17± 0.11	96.76± 0.13	92.25±0.77	**98.86±0.43**	98.47±0.22
MNIST	90.26± 0.62	93.38± 0.82	92.54± 0.76	88.35±1.86	95.47±1.64	**97.69±0.95**
Yale	69.59± 5.26	72.71± 5.54	74.28± 4.64	70.89±5.25	82.76±4.97	**84.89±5.32**
PIE	85.38 ± 0.13	87.29± 0.19	89.92± 0.16	82.15±0.87	90.34±0.83	**93.15±0.11**
ORL	91.49± 1.84	91.45± 1.27	92.41± 1.07	89.19±0.52	95.26±0.93	**96.19±0.89**
COIL-20	92.57± 0.89	91.33± 0.35	94.96± 0.27	90.47±0.94	96.58±0.34	**97.63±0.15**

Table 3: The learned weights and classification accuracies of KFDA with the single kernel on Ionosphere.

σ	Weights learned by MKL-TR	Weights learned by MKL-MFA	KFDA with the single kernel
0.10	0.02	0.01	64.53
0.22	0.02	0.01	65.75
0.46	0.06	0.03	90.82
1.00	0.36	0.40	92.24
2.15	0.37	0.42	93.07
4.46	0.13	0.11	91.98
10.00	0.03	0.02	90.17
21.54	0.02	0.01	88.67
46.42	0.001	0.001	87.73
100.00	0.01	0.002	85.26

Table 4: 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

using MKL-MFA, as displayed in Figure 1. It is shown that the projected results of MKL-MFA and MKL-SRTR have the better separability than those of MKL-DR and MKL-TR. The separability of embedded data of MKL-MFA is much clearer than that of other models. Consequently, our model is also superior to others on high-dimensional data.

To further evaluate the effectiveness of MKL-MFA, we used bearing vibration signals of accelerometer sensors under different operating loads as a real world dataset. The vibration signals were collected by using a 16 channel digital audio tape (DAT) recorder at the sampling frequency 12 kHz. The experimental vibration data were divided into four datasets, named as D_IRF, D_ORF, D_BF, and D_MIX shown in Table 4, where "07", "14", "21", and "28" mean that fault diameters are 0.007, 0.014, 0.021, and 0.028 inches, respectively [24]. Signals were selected randomly to form training and testing sets with ratio 1:1.

Firstly, vibration signals were transformed into 10 time domain features, 3 frequency domain features, and 16 time-frequency domain features [24]. Secondly, deferent dimensionality algorithms were carried out to extract low dimensional features from the transformed signals. Finally, we used SVM to train and test low dimensional features to compare our method with other DR methods. The classification accuracy rates are reported in Table 5. It can be seen that, compared with other algorithms, MKL-MFA achieves much better performance on all datasets, which further validates the effectiveness of MKL-MFA for feature extraction of vibration signals in real applications.

5. Conclusions

In this paper, we extend the Marginal Fisher Criterion to the multiple kernel case. Based on the extended criterion, a new

Table 5: The classification accuracy rates on four bearing vibration signal datasets.

Datasets	MKL-TR	MKL-DR	MKL-SRTR	EMFA	MKL-MFA
D_MIX	0.9363	0.9257	0.9485	0.9854	**0.9914**
D_IRF	0.9415	0.9151	0.9312	0.9627	**0.9862**
D_ORF	0.9228	0.9238	0.9554	0.9716	**0.9848**
D_BF	0.9086	0.8992	0.9027	0.9285	**0.9537**

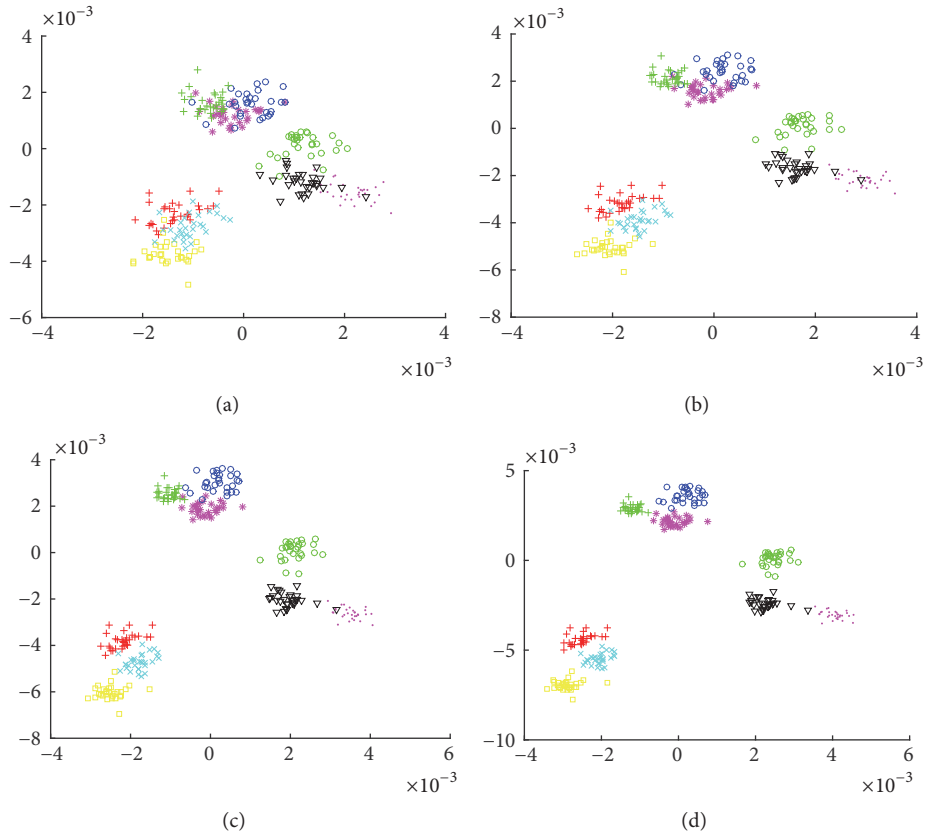

Figure 1: *Comparison of two-dimensional embedded results obtained by different algorithms on the first 9 classes of PIE.* (a) Projection of training data using supervised MKL-DR. (b) Projection of training data using supervised MKL-TR. (c) Projection of training data using supervised MKL-SRTR. (d) Projection of training data using MKL-MFA.

multiple kernel-based dimensionality reduction algorithm termed as MKL-MFA is proposed for supervised nonlinear dimensionality reduction. Without prior information on data distributions, MKL-MFA is more general for multiple kernel discriminant analysis. Experimental results on benchmark and real world datasets validate the promising performance of MKL-MFA, respectively. In the near future, we intend to improve our model by introducing deep kernel networks and study nonlinear dimensionality reduction methods via deep models.

Data Availability

The benchmark data used to support the findings of this study have been deposited in the UCI and face image repository (http://archive.ics.uci.edu/ml/datasets.html, http://www.face-rec.org/databases/, http://www.ri.cmu.edu/projects/project_418.html, http://web.mit.edu/emeyers/www/face_databases.html, https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html, http://yann.lecun.com/exdb/mnist/, http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php). The bearing vibration signals of accelerometer sensors under different operating loads, provided by the Bearing Data Center of the Case Western Reserve University, have been validated in many research works and become a standard dataset for bearing studies (http://csegroups.case.edu/bearingdatacenter/home).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by Natural Science Foundation of Jiangsu Province [grant numbers BK20170273 and BK20180174] and the National Natural Science Foundation of China [grant numbers 61801198, 41672324, and 41704115].

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 and 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] S. Yan, D. Xu, B. Zhang, H. 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.

[8] 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.

[9] 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, 2009.

[10] B. Liu, Y. Zhou, Z. Xia, P. Liu, Q. Yan, and H. Xu, "Spectral regression based marginal Fisher analysis dimensionality reduction algorithm," *Neurocomputing*, vol. 277, no. 14, pp. 101–107, 2018.

[11] 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.

[12] 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.

[13] 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.

[14] 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 and Learning Systems*, vol. 22, no. 11, pp. 1796–1808, 2011.

[15] V. Raviteja, V. P. Boda, and R. Chellappa, "MKL-RT: Multiple Kernel Learning for Ratio-trace Problems via Convex Optimization," Eprint Arxiv (2014).

[16] R. Hettich and K. O. Kortanek, "Semi-infinite programming: theory, methods, and applications," *SIAM Review*, vol. 35, no. 3, pp. 380–429, 1993.

[17] 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.

[18] 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.

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

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

[21] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang, "Trace Ratio vs. Ratio Trace for Dimensionality Reduction," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR '07)*, pp. 1–8, 2007.

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

[23] D. Cai, X. He, and J. Han, "SRDA: an Efficient Algorithm for Large Scale Discriminant Analysis," in *Proceedings of the IEEE Trans. Knowl. Data Eng*, vol. 20, pp. 1–12, 2008.

[24] 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.