Knowl Inf Syst (2011) 27:473–490
DOI 10.1007/s10115-010-0306-z

SHORT PAPER

Locality preserving multimodal discriminative learning for supervised feature selection

Zhao Zhang · Ning Ye

Received: 6 August 2009 / Revised: 22 March 2010 / Accepted: 20 May 2010 /
Published online: 10 June 2010

Abstract Feature selection has been an important preprocessing step in high-dimensional data analysis and pattern recognition. In this paper, we propose a locality preserving multimodal discriminative learning method called LPMDL for supervised feature selection, which arises by solving two standard eigenvalue problems and seeks to find a pair of optimal transformations for two sets of multivariate data in different classes. This topic can optimally discover the local structure information of the given data hided in the original space and aims at structuring an effective low-dimensional embedding space, under which LPMDL keeps nearby data pairs in the same class close and between-class data pairs apart, and the projections of the original data in different classes can be appropriately separated from each other. LPMDL can be performed either in the input space or the reproducing kernel Hilbert space which gives rise to the kernelized version of LPMDL. We also evaluate the feasibility and efficiency of the LPMDL approach by conducting extensive data visualization and classification tasks. Experimental results on a broad range of data sets show LPMDL tends to capture the intrinsic structure characteristics of the samples data due to the effective representation of the points and achieves similar or even better performance than the conventional PCA, NPE, LPP and LFDA methods.

Keywords Locality preservation · Feature selection · Data visualization · Within-class multimodality · Classification · Discriminative learning

Z. Zhang (✉) · N. Ye
School of Information Science and Technology, Nanjing Forestry University,
210037 Nanjing, China
e-mail: cszzhang@gmail.com

N. Ye
School of Computer Science and Technology, Shandong University,
250100 Jinan, China
e-mail: ye.ning@yahoo.com.cn

1 Introduction

More and more scientific research and real-word applications require to deal with the high-dimensional image data, which leads to a deep study of the feature selection approaches [8,7,13,29,19,16,6,23,26,28]. Feature selection has been an important preprocessing step in pattern recognition and high-dimensional data analysis, which attracts much attention and leads us to consider the problems of feature selection that allow one to represent the data in a reduced pace, while most of intrinsic information hided in the data can be effectively preserved. Once feature selection methods are performed appropriately, then we can utilize the low-dimensional representations of data for a variety of succeeding tasks, such as visualization, classification and image recognition.

Based on whether the class labels and constraints are adopted, feature selection methods can be divided into supervised methods [8] that evaluates feature relevance by the correlation between the features and the class labels or the constraints, and unsupervised cases [7,6], evaluating feature relevance by the capability of keeping the local structure preserving ability. To embed the data well, it is essential and important to preserve the spatial local structure of the points. Mapping data from a high-dimensional space into a low-dimensional embedding space is considered to be locality preserving if points nearby in the original space are still compact in the embedding space [17]. The local structures between the samples data can be regarded as the spatial distribution or location of data in the original input space and the learnt feature space [32]. Locality preserving projection (LPP) [10] keeps nearby data pairs in the input space close in the found embedding space, by which the multimodal data can be embedded without losing its intrinsic structures. However, Sugiyama [24] has pointed that LPP tends to make samples of different classes overlapped if they are close in the input space. Locality pursuit embedding (LPE) [16] has been introduced to consider and preserve the locality variation information. Roweis and Saul have proposed the locally linear embedding (LLE) [19], which assumes that any one datum could be reconstructed by using its local nearest neighbors in the original space and this local reconstruction relationship can still hold in the obtained low-dimensional space. Neighbor-hood preserving embedding (NPE) [11] is the linear approximation to the LLE method [19] and aims at preserving the global Euclidean structure and aims to preserve the local neighborhood structure of the data manifold. It is worthy of noticing that LPP and NPE do not take into account the class information of samples which is actually useful in machine learning and are developed originally for the unsupervised cases. Alternatively, though these methods can be extended to the supervised learning fields when the class information is used to construct the weights, it does not necessarily work effectively in the supervised learning scenarios. Sugiyama has taken the local structure information and the class information of the samples into account and proposed the local Fisher discriminant analysis (LFDA) [24] for supervised feature selection or, namely, dimensionality reduction. Numerical results show that LFDA tends to achieve with-class compactness and between-class separability.

In recent years, the locality preservation–based learning approaches are significantly studied and developed for feature selection and pattern recognition. These successful applications of the locality-based methods [10,24,27,11,13,19,16,12] have inspired us to pay much more attention to the locality-based techniques. In this paper, we propose an effective algorithm for supervised feature selection, which we refer to as locality preserving multimodal discriminative learning (LPMDL), setting a graph incorporating the neighborhood information of samples and aiming to compute a pair of bases ω_x and ω_y for two data sets in different classes. LPMDL can achieve between-class separability and preserve the local structure

and within-class multimodal structure of the given data simultaneously. This algorithm is interesting from some remarkable perspectives:

1. LPMDL is a linear method, which makes it applicable for practical applications, and similar to LFDA [24], it does not project the multimodal data in the same class to a single cluster.
2. For feature selection, the embedding transformation can be effectively computed and be simply applied to any new point to locate it in the reduced space. Moreover, LPMDL can resolve the XOR problems and the results of the XOR data set are superior to those of some established feature selection approaches, such as, principal component analysis (PCA) [9],LPP [10], NPE [11] and LFDA [24].
3. LPMDL may be conducted either in the original space or in the reproducing kernel Hilbert space (RKHS) Thus, LPMDL may be extended to the nonlinear scenarios by employing the so-called kernel trick [20].
4. For visualization, the projections of the points in different classes can be effectively separated from each other in the feature space obtained by LPMDL. Moreover, LPMDL aims to preserve the local structure of the data, therefore, it is likely that a local neighbor search in the input space will yield the similar results in the low-dimensional embedding space.
5. By introducing the unlabeled samples for learning, LPMDL can be extended to the semi-supervised case. The detailed implementation of the semi-supervised feature selection method is beyond the focus of this paper and we will discuss it in another paper.

As a result of all these perspectives, we expect that the proposed feature selection technique to be widely used in the fields of data visualization and pattern recognition. The rest of the paper is organized as follows. In Sect. 2, we present a sensitivity analysis of the proposed algorithm and show its fundamental properties. In Sect. 3, we numerically evaluate the performance of our method and some existing feature selection methods by visualization analysis and classification using some benchmark UCI and real-world data sets. Finally, we conclude this paper and raise some issues for future works in Sect. 4.

2 Locality preserving multimodal discriminative learning

2.1 Basic idea

In Figs. 1 and 2, the dimensionality reduction results obtained by PCA [9], LPP [10], NPE [11], LFDA [24] and our LPMDL method on the *Toy* and *XOR* data sets and the distributions of the original data are shown, where two-dimensional two-class samples are embedded into a one-dimensional space. The line in each figure denotes the one-dimensional embedding space, on which the data points are projected, found by these methods. In LPP, the affinity matrix A is determined by the heat kernel method [2]. For the simplest data set shown in Fig. 1, PCA, LPP and LFDA can find the better embedding spaces and give the promising results where samples of different classes are nicely separated. As illustrated in Fig. 1, NPE performs worse than the other methods if samples in a class form separate clusters, i.e. *multimodal*. For the *XOR* data set in Fig. 2, no matter unsupervised PCA, LPP, NPE methods or supervised LFDA method cannot perform effectively on the data set due to their obtained one-dimensional embedding space that the samples data are projected on. In fact, the *XOR* data set has within-class multimodality in each class, which makes only one set of optimal transformations cannot embed the data points appropriately. We see from Fig. 2 that PCA,

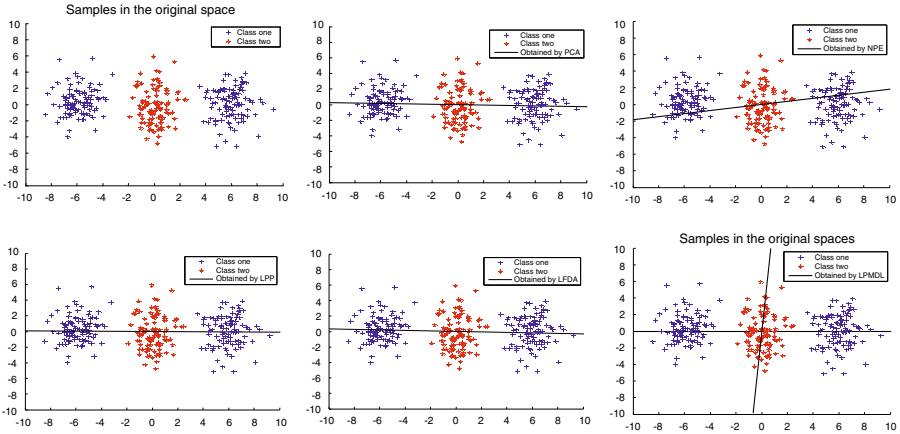


Fig. 1 Examples of feature selection by PCA, NPE, LPP, LFDA, LPMDL on the *Toy* data set

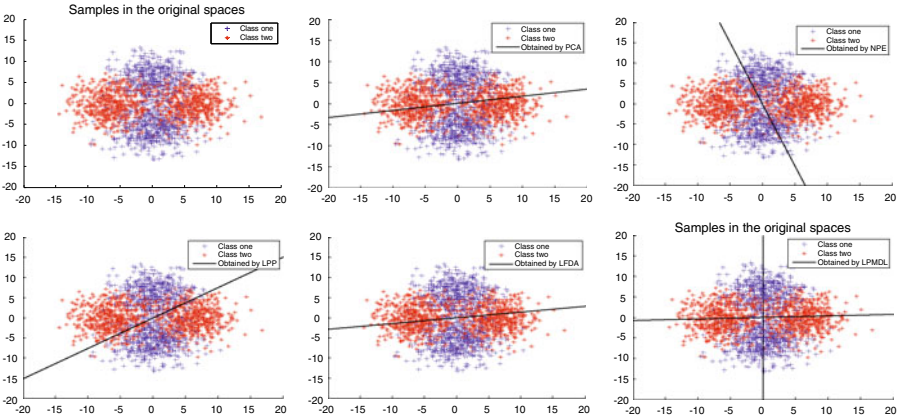


Fig. 2 Examples of feature selection by PCA, NPE, LPP, LFDA, LPMDL on the *XOR* data set

LPP, NPE and LFDA tend to mix the projections of the multimodal data of different classes when they are compact in the original input space. In order to embed the multimodal data well, we utilize the class labels of samples and propose LPMDL to perform dimensionality reduction on the data, whose aim is to compute two sets of optimal projection transformations for two sets of multivariate data in different classes. More importantly, we evaluate the levels of the between-class spread or scatter and the within-class spread or scatter in a local manner, allowing us to achieve between-class separation and within-class multimodal structure preservation and local structure preservation at the same time.

2.2 Learning linear LPMDL

In this section, we formulate the supervised locality preservation–guided multimodal discriminative learning for feature selection as the following. Given two sets of labeled multivariate data $X, Y \in \mathbb{R}^n$ and let X, Y be $X = (x_0, x_1, \ldots, x_{m_x-1})$ and $Y = (y_0, y_1, \ldots, y_{m_y-1})$ with two different class labels. The objective is to seek a pair of projection transformations, ω_x and ω_y, and the transformations map the samples data onto a set of points in a

low-dimensional embedding space. To improve the tightness among the similar patterns in the same class and separate the dissimilar patterns in different classes better, we consider shrinking the distances between similar patterns by minimizing $\sum_{i=1}^{m_x}\sum_{j=1}^{m_x} A_{xx}||\omega_x^T x_i - \omega_x^T x_j||^2$ $\left(\sum_{i=1}^{m_y}\sum_{j=1}^{m_y} A_{yy}||\omega_y^T y_i - \omega_y^T y_j||^2\right)$, while expanding dissimilar ones by maximizing $\sum_{i=1}^{m_x}\sum_{j=1}^{m_y} A_{xy}||\omega_y^T x_i - \omega_y^T y_j||^2$ $\left(\sum_{i=1}^{m_y}\sum_{j=1}^{m_x} A_{yx}||\omega_y^T y_i - \omega_y^T x_j||^2\right)$, where the transformations $\omega_x^T X$ and $\omega_y^T Y$ are, respectively, the low-dimensional representations of the X and Y sets and the weights A_{xx}, A_{yy}, A_{xy} and A_{yx} are introduced for preserving the locality.

Let G represent a graph with m nodes, then we put an edge between nodes i and j if sample $x_i(y_i)$ and sample $x_j(y_j)$ are local neighbors. Furthermore, the local neighbors of sample $x_i(y_i)$ may be defined as the following two popular approaches [10,1]:

(1) ε-hypersphere measure: if $||x_i - x_j|| \le \varepsilon(||y_i - y_j|| \le \varepsilon)$, thus we say that sample $x_j(y_j)$ is the local neighbor of sample $x_i(y_i)$, where $\varepsilon \in \mathbb{R}$ is a user specified control parameter.

(2) k-nearest neighborhood measure: if sample $x_j(y_j)$ is among the k-nearest neighbors of sample $x_i(y_i)$, thus we say sample $x_j(y_j)$ is the local neighbor of sample $x_i(y_i)$, where the parameter, $k \in \mathbb{N}$.

Once the graph G is constructed, on the basis of the definitions of local neighborhood, we can define the similarity matrices $A^x = \left\{A_{ij}^x\right\}_{i,j}^{m_x}$ and $A^y = \left\{A_{ij}^y\right\}_{i,j}^{m_y}$ as the following, where

$$A_{ij}^x = \begin{cases} \exp\left(-||x_i - x_j||^2/\kappa\right), & \text{if sample } x_i \text{ is the local neighbor of sample } x_j \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

$$A_{ij}^y = \begin{cases} \exp\left(-||y_i - y_j||^2/\kappa\right), & \text{if sample } y_i \text{is the local neighbor of sample } y_j \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where the tunable parameter κ is a positive scalar and $||.||$ denotes the L2-norm with Euclidean metric. In fact, Eqs. 1 and 2 can reflect the locality around each data point, that is, the smaller the distances $||x_i - x_j||^2$ $(||y_i - y_j||^2)$, the closer the samples data, and thus the larger $A_{ij}^x(A_{ij}^y)$. For dimensionality reduction, local neighbors will have similar embeddings, so the data points lying on a dense area are likely to have the same label [33]. Thus, if x_j and x_i are in the same class, we say they are local neighbors. Let A_{xx}, $A_{yy}$and A_{xy} be the matrices with $A_{xx} = R_{xx} - A^x \times A^x$, $A_{yy} = R_{yy} - A^y \times A^y$ and $A_{xy} = R_{xy} - A^x \times A^y$. Let A_{ij} denote the (i,j)th entry of A, thus $R_{xx}(R_{yy},R_{xy})$ is a diagonal matrix whose entries are column (or row due to the symmetry) sum of the matrix $A^x \times A^x(A^y \times A^y,A^x \times A^y)$, i.e., $R_{xx} = \sum_i (A^x \times A^x)_{i,j}$, and the similar expressions exist for R_{yy} and R_{xy}. In the computations, we compute the dot products between two matrices, i.e., for matrices C, D with the same size, $(C \times D)_{ij} = C_{ij}D_{ij}$. Since local neighbors tend to have the similar embeddings on the data manifold, the points distributed in a dense area commonly have the same label, thus if sample $x_j(y_j)$ has the same label with sample $x_i(y_i)$, we say they are mutually local neighbors.

Noticing that $\omega^T x$ means that transformation ω basically projects the points to a set of useful features in the reduced feature space [25], in which samples of different classes can be easily partitioned from each other. Ideally, the feature set should be as compact as possible that means that the small rank for $\omega^T\omega$ or ω, since $\text{rank}(\omega^T\omega) = \text{rank}(\omega)$ will be desired. Here, we aim to minimize it by finding an eigen-decomposition Tr of $\omega^T\omega$, that is, $Tr(\omega^T\omega) = V\Lambda V^T$, thus $\text{rank}(\omega^T\omega) = \text{rank}(\Lambda) = ||\Lambda||_0$, but a direct optimization of

the zero norm is not practical to deal with. Therefore we approximate it by the Euclidean (L2-) norm $||\Lambda||_2 = ||\omega||_2$ in the computations. Then we formulate the following objective functions for discriminative learning:

$$\underset{\omega_x}{Min} \frac{\frac{1}{2}||\omega_x||^2 + \frac{M_{xx}}{2}\sum_{i=1}^{m_x}\sum_{j=1}^{m_x} A_{xx}||\omega_x^T x_i - \omega_x^T x_j||^2}{\sum_{i=1}^{m_x}\sum_{j=1}^{m_y} A_{xy}||\omega_x^T x_i - \omega_x^T y_j||^2} \tag{3}$$

$$\underset{\omega_y}{Min} \frac{\frac{1}{2}||\omega_y||^2 + \frac{M_{yy}}{2}\sum_{i=1}^{m_y}\sum_{j=1}^{m_y} A_{yy}||\omega_y^T y_i - \omega_y^T y_j||^2}{\sum_{i=1}^{m_y}\sum_{j=1}^{m_x} A_{yx}||\omega_y^T y_i - \omega_y^T x_j||^2} \tag{4}$$

where T denotes the transpose of a matrix and m_x and m_y are the numbers of samples in class 1 and class 2, respectively. In the experiments, we consistently need $m_x = m_y$. Here, we add two scaling parameters M_{xx} and M_{yy} to balance the contributions of two terms in the numerators of Eqs. 3 and 4. The intuitions behind Eqs. 3 and 4 are brief and natural. We aim to select the canonical features with better locality preserving ability. More specifically, if there is the same class label between two samples, a canonical feature should be the one on which those two samples are close to each other; on the other hand, if the samples have different class labels, a canonical feature should be the one on which those two samples are far away from each other. Moreover, Eqs. 3 and 4 realize the feature selection approach according to the features' locality preserving abilities.

Theorem 1 *The optimal transformation ω_x can be solved by computing the eigenvectors according to the first d smallest eigenvalues of the matrix $\left(I + M_{xx}A^{(XX)} - M_{xy}A^{(XY)}\right)$, where I is the identity matrix.*

Justification of Theorem 1. From Eq. 3, we expect the distances among data samples in the same class to be as small as possible and to be as away from the samples with different class labels as possible. Besides, it is also worthy of noting the singularity. In order to avoid involving the matrix inverse operation and to ensure the computational stability, here we reformulate the objective function in Eq. 3 to the following:

$$\underset{\omega_x}{Min}\ \frac{1}{2}||\omega_x||^2 + \frac{M_{xx}}{2}\omega_x^T A^{(XX)}\omega_x - \frac{M_{xy}}{2}\omega_x^T A^{(XY)}\omega_x \tag{5}$$

with respect to $\omega_x^T\omega_x = I$ and I is the identity matrix. Here, we add another scaling parameter M_{xy} to balance the contributions of three terms in Eq. 5. Intuitively, the distances involved among samples in the same should typically be close to the expected metric, thus we empirically set $M_{xx} = 1$ and $M_{xy} > 1$, respectively for optimization. Here, we formulate $A^{(XX)}$ and $A^{(XY)}$ in a matrix form as follows:

$$\begin{aligned}
A^{(XX)} &= \sum_{i=1}^{m_x}\sum_{j=1}^{m_x}(A_{xx})_{i,j}\,||x_i - x_j||^2 \\
&= \sum_{i=1}^{m_x}\left(\sum_{j=1}^{m_x}(A_{xx})_{i,j}\right)x_i x_i^T + \sum_{j=1}^{m_x}\left(\sum_{i=1}^{m_x}(A_{xx})_{i,j}\right)x_j x_j^T - 2\sum_{i,j=1}^{m_x}(A_{xx})_{i,j}\,x_i x_j^T, \\
&= 2\sum_{i=1}^{m_x} D_{ii}x_i x_i^T - 2XA_{xx}X^T \\
&= 2XF_{xx}X^T
\end{aligned} \tag{6}$$

$$\begin{aligned} A^{(XY)} &= \sum_{i=1}^{m_x}\sum_{j=1}^{m_y}\left(A_{xy}\right)_{i,j}||x_i - y_j||^2 \\ &= \sum_{i=1}^{m_x} W_{ii}x_i x_i^T + \sum_{j=1}^{m_y} M_{jj}y_j y_j^T - 2XA_{xy}Y^T, \\ &= XWX^T + YMY^T - 2XA_{xy}Y^T \end{aligned} \tag{7}$$

where $(A_{xx})_{i,j}$ is the (i, j)th entry of A_{xx} and $D(W,M)$ is a diagonal matrix and its ith entry equals the sum of the entries in the ith row (or the ith column due to the symmetry) of the matrix $A_{xx}(A_{xy})$, i.e.

$$D_{ii} = \sum_{j=1}^{m_x}(A_{xx})_{i,j}, \quad W_{ii} = \sum_{j=1}^{m_x}\left(A_{xy}\right)_{i,j}, \quad M_{jj} = \sum_{i=1}^{m_y}\left(A_{xy}\right)_{i,j}. \tag{8}$$

Forming the Lagrangian of Eq. 5 with the multipliers λ_x, then we can obtain

$$L(\omega_x, \lambda_x) = \frac{1}{2}||\omega_x||^2 + \frac{M_{xx}}{2}\omega_x^T A^{(XX)}\omega_x - \frac{M_{xy}}{2}\omega_x^T A^{(XY)}\omega_x - \frac{\lambda_x}{2}(\omega_x^T\omega_x - I). \tag{9}$$

By computing $\partial L(\omega_x, \lambda_x)/\partial\omega_x = 0$, we get the following eigenvalue problem:

$$\left(I + M_{xx}A^{(XX)} - M_{xy}A^{(XY)}\right)\omega_x = \lambda_x\omega_x. \tag{10}$$

Clearly, it is a scale-reduced typical eigenvalue problem, from which the optimal projection transformation $\omega_x = \left(\omega_{x_{[1]}}|\omega_{x_{[2]}}|\ldots|\omega_{x_{[d]}}\right), d \leq n$ that minimizes the objective function in Eq. 5 can be effectively obtained by solving the eigenvectors of $\left(I + M_{xx}A^{(XX)} - M_{xy}A^{(XY)}\right)$ corresponding to the first d smallest eigenvalues.

Theorem 2 *The optimal transformation ω_y can be solved by computing the eigenvectors according to the first d smallest eigenvalues of the matrix $\left(I + M_{yy}A^{(YY)} - M_{xy}A^{(XY)}\right)$, where I is the identity matrix.*

Justification of Theorem 2. Let $A_{yx} = R_{yx} - A^y \times A^x$, thus $A^{(YX)} = \sum_{i=1}^{m_y}\sum_{j=1}^{m_x}\left(A_{yx}\right)_{i,j}$ $||y_i - x_j||^2 = YMY^T + XWX^T - 2XA_{xy}Y^T = A^{(XY)}$ and $A_{yx} = A_{xy}$. Analogous to the computations of the optimal ω_x, we reformulate the objective in function in Eq. 4 to the following problem with respect to $\omega_y^T\omega_y = I$:

$$\underset{\omega_y}{Min}\ \frac{1}{2}||\omega_y||^2 + \frac{M_{yy}}{2}\omega_y^T A^{(YY)}\omega_y - \frac{M_{xy}}{2}\omega_y^T A^{(XY)}\omega_y, \tag{11}$$

Similarly, the matrix $A^{(YY)}$ can be interpreted in a matrix form as the following:

$$\begin{aligned} A^{(YY)} &= \sum_{i=1}^{m_y}\sum_{j=1}^{m_y}\left(A_{yy}\right)_{i,j}||y_i - y_j||^2 \\ &= \sum_{i=1}^{m_y}\left(\sum_{j=1}^{m_y}\left(A_{yy}\right)_{i,j}\right)y_i y_i^T + \sum_{j=1}^{m_y}\left(\sum_{i=1}^{m_y}\left(A_{yy}\right)_{i,j}\right)y_j y_j^T - 2\sum_{i,j=1}^{m_y}\left(A_{yy}\right)_{i,j}y_i y_j^T, \\ &= 2\sum_{i=1}^{m_y} Z_{ii}y_i y_i^T - 2YA_{yy}Y^T \\ &= 2YF_{yy}Y^T \end{aligned} \tag{12}$$

where Z is a m_y-dimensional diagonal matrix with ith diagonal input element being

$$Z_{ii} = \sum_{j=1}^{m_y} \left(A_{yy}\right)_{i,j}. \tag{13}$$

In order to ensure that distances of the points involved among data samples with the same class labels should be typically close to the expected metric, we empirically set $M_{yy} = 1$ and $M_{xy} > 1$ as well. Thus, forming the Lagrangian of Eq. 11 with the multipliers λ_y and by zeroing it, we obtain the following similar scale-reduced typical eigenvalue problem:

$$\left(I + M_{yy}A^{(YY)} - M_{xy}A^{(XY)}\right)\omega_y = \lambda_y\omega_y, \tag{14}$$

from which $\omega_y = \left(\omega_{y_{[1]}}|\omega_{y_{[2]}}|\ldots|\omega_{y_{[d]}}\right)$ can be effectively obtained by computing the eigenvectors of the matrix $\left(I + M_{yy}A^{(YY)} - M_{xy}A^{(XY)}\right)$ corresponding to the first d smallest eigenvalues.

Computationally, the LPMDL approach is finally transformed to two standard eigenvalue problems based on three symmetric matrices, so LPMDL can be easily computed by the eigen-decompositions. Once the pairs of the optimal projection transformations $\left(\omega_x, \omega_y\right)$ are obtained, dimensionality reduction can be performed in the forms of $\omega_x^T X$ and $\omega_y^T Y$. In the experiments, we keep the local information A_{ij}^x and A_{ij}^y unchanged and attempt to ensure that if points are close in the original input space, then after dimensionality reduction using our LPMDL method, the points in the found embedding space are still compact with each other, which will be validated by data visualization and classification experiments.

2.3 Kernel generalization

Till now, we only focused on linear feature selection. Next, we will extend our discussion further to nonlinear feature selection scenarios by the so-called kernel trick [20] and propose the kernelized approach, which we refer to as KLPMDL in further readings. Implicitly in the kernel Hilbert space connected to the kernel function K that is used. According to [31], a kernel is a function in the input space and simultaneously is the inner product in the embedding space through the kernel-induced measure-based nonlinear mapping. More specifically, a kernel can be formulated as the dot-product form of $K\left(x_i, x_j\right) = \langle\phi\left(x_i\right), \phi\left(x_j\right)\rangle = \left(\phi\left(x_i\right)\right)^T \phi\left(x_j\right)$. Since for each kernel function, there exists a mapping ϕ corresponds to a scalar product and maps input data x to $\phi(x)$, here we define the following mapping $\phi : \mathbb{R}^n \to \mathrm{H}^p, (p > n), \quad \phi(x) = K(., x)$, i.e., ϕ maps each data x_i to the function $K(., x_i)$, where the first argument of K is free and the second is fixed to x_i.

Mika et al. have proved that every solution $\beta \in \mathrm{H}$ (usually high-dimensional) in the kernel feature space can be written as an expansion in terms of the mapped training data [15]. Therefore, the projection vectors ω_x^ϕ and ω_y^ϕ in the high-dimensional kernel space can be rewritten as the following:

$$\omega_x^\phi = \sum_{i=1}^{m_x} \alpha_x^i \phi\left(x_i\right) = \phi(X)\alpha_x, \omega_y^\phi = \sum_{i=1}^{m_y} \alpha_y^i \phi\left(y_i\right) = \phi(Y)\alpha_y. \tag{15}$$

By substituting the basis vectors ω_x^ϕ and ω_y^ϕ into the original optimization problems, we can formulate KLPMDL as follows. Let $\phi(X) = \left(\phi\left(x_1\right), \phi\left(x_2\right), \ldots, \phi\left(x_{m_x}\right)\right)$,$\phi(Y) = \left(\phi\left(y_1\right), \phi\left(y_2\right), \ldots, \phi\left(y_{m_y}\right)\right)$ and column vectors $\alpha = \left[\alpha_x, \alpha_y\right] = \left[\left(\alpha_{x_{[1]}}|\alpha_{x_{[2]}}|\ldots|\alpha_{x_{[d]}}\right),\right.$

$\left(\alpha_{y_{[1]}}|\alpha_{y_{[2]}}|\ldots|\alpha_{y_{[d]}}\right)\Big]$ be $d$pairs of projection vectors in the kernel space. Therefore, Eq. 5 can be rewritten as

$$\begin{aligned} &\underset{\omega_x^\phi}{Min}\ \frac{1}{2}||\omega_x^\phi||^2 + \frac{M_{xx}}{2}\left(\omega_x^T \Re \omega_x\right)^\phi - \frac{M_{xy}}{2}\left(\omega_x^T \aleph \omega_x\right)^\phi \\ &S.t.\left(\omega_x^T \omega_x\right)^\phi = I, \end{aligned} \tag{16}$$

where matrices $A^{(XX)}$ and $A^{(XY)}$ are formulated as

$$A^{(XX)} = \sum_{i,j=1}^{m_x} \left(A_{xx}^\phi\right)_{i,j} \left(\phi\left(x_i\right) - \phi\left(x_j\right)\right)\left(\phi\left(x_i\right) - \phi\left(x_j\right)\right)^T, \tag{17}$$

$$A^{(XY)} = \sum_{i=1}^{m_x}\sum_{j=1}^{m_y} \left(A_{xy}^\phi\right)_{i,j} \left(\phi\left(x_i\right) - \phi\left(y_j\right)\right)\left(\phi\left(x_i\right) - \phi\left(y_j\right)\right)^T, \tag{18}$$

where matrices A_{xx}^ϕ, A_{xy}^ϕ are defined for preserving the local relations among samples in the kernel space. Let $S(Q,L)$ be a diagonal matrix and its ith diagonal entry equals the sum of the entries in the ith row (or the ith column due to the symmetry) of $A_{xx}^\phi(A_{xy}^\phi)$, that is, $S_{ii} = \sum_j \left(A_{xx}^\phi\right)_{i,j}$, $Q_{ii} = \sum_j \left(A_{xy}^\phi\right)_{i,j}$ and $L_{jj} = \sum_i \left(A_{xy}^\phi\right)_{i,j}$. Therefore, we can compute the Laplacian matrix [5] over A_{xx}^ϕ as $E = S - A_{xx}^\phi$. By computing $\left(\omega_x^T A^{(XX)}\omega_x\right)^\phi$, we get

$$\begin{aligned} \left(\omega_x^T A^{(XX)}\omega_x\right)^\phi &= \alpha_x^T\left(\phi\left(x\right)\right)^T \left(\sum_{i,j=1}^{m_x}\left(A_{xx}^\phi\right)_{i,j}\left(\phi\left(x_i\right) - \phi\left(x_j\right)\right)\left(\phi\left(x_i\right) - \phi\left(x_j\right)\right)^T\right)\phi\left(x\right)\alpha_x \\ &= \alpha_x^T\left(\phi\left(x\right)\right)^T \left(\sum_{i=1}^{m_x}\left(\sum_{j=1}^{m_x}\left(A_{xx}^\phi\right)_{i,j}\right)\phi\left(x_i\right)\left(\phi\left(x_i\right)\right)^T\right. \\ &\quad \left. + \sum_{j=1}^{m_x}\left(\sum_{i=1}^{m_x}\left(A_{xx}^\phi\right)_{i,j}\right)\phi\left(x_j\right)\left(\phi\left(x_j\right)\right)^T\right)\phi\left(x\right)\alpha_x \\ &\quad -\alpha_x^T\left(\phi\left(x\right)\right)^T\left(\sum_{i,j=1}^{m_x}\left(A_{xx}^\phi\right)_{i,j}\phi\left(x_i\right)\left(\phi\left(x_j\right)\right)^T\right. \\ &\quad \left. + \sum_{i,j=1}^{m_x}\left(A_{xx}^\phi\right)_{i,j}\phi\left(x_j\right)\left(\phi\left(x_i\right)\right)^T\right)\phi\left(x\right)\alpha_x \\ &= 2\alpha_x^T\sum_{i=1}^{m_x}\left(\phi\left(x_i\right)\right)^T\phi\left(x_i\right)S_{ii}\left(\phi\left(x_i\right)\right)^T\phi\left(x_i\right)\alpha_x \\ &\quad -2\alpha_x^T\left(\phi\left(x\right)\right)^T\phi(x)A_{xx}^\phi\left(\phi\left(x\right)\right)^T\phi\left(x\right)\alpha_x \\ &= 2\alpha_x^T E\alpha_x. \end{aligned} \tag{19}$$

Similarly, let matrix V be $V = \left(K_{xx} Q K_{xx} + K_{xy} L K_{xy}^T - 2K_{xx} A_{xy}^{\phi} K_{xy}^T\right)$, thus we can obtain

$$\begin{aligned}\left(\omega_x^T A^{(XY)} \omega_x\right)^{\phi} &= \alpha_x^T \left(\phi(x)\right)^T \left(\sum_{i=1}^{m_x}\sum_{j=1}^{m_y} \left(A_{xy}^{\phi}\right)_{i,j} \left(\phi(x_i) - \phi\left(y_j\right)\right)\left(\phi(x_i) - \phi\left(y_j\right)\right)^T\right) \phi(x)\alpha_x \\ &= \alpha_x^T \left(K_{xx} Q K_{xx} + K_{xy} L K_{xy}^T - 2K_{xx} A_{xy}^{\phi} K_{xy}^T\right)\alpha_x \\ &= \alpha_x^T V \alpha_x \end{aligned} \tag{20}$$

with $K_{xx} = \left(\phi(X)\right)^T \phi(X)$, where K_{xx} is the kernel matrix among the data samples in X set and K_{xy} is the kernel matrix among the data samples in X and Y sets, which is represented as $K_{xy} = \left(\phi(X)\right)^T \phi(Y)$. Let $U = 2E$, and by substituting Eqs. 19 and 20 into problem of Eq. 16, we can rewrite Eq. 16 as

$$\begin{aligned} &\underset{\alpha_x}{Min}\ \frac{1}{2}\alpha_x^T K_{xx}\alpha_x + \frac{M_{xx}}{2}\alpha_x^T U \alpha_x - \frac{M_{xy}}{2}\alpha_x^T V \alpha_x \\ &S.t.\ \alpha_x^T K_{xx}\alpha_x = I. \end{aligned} \tag{21}$$

Forming the Lagrangian of Eq. 21 with the multipliers λ_x, we obtain the following formulation:

$$L(\alpha_x, \lambda_x) = \frac{1}{2}\alpha_x^T K_{xx}\alpha_x + \frac{M_{xx}}{2}\alpha_x^T U \alpha_x - \frac{M_{xy}}{2}\alpha_x^T V \alpha_x - \frac{\lambda_x}{2}(\alpha_x^T K_{xx}\alpha_x - I). \tag{22}$$

By computing $\partial L(\alpha_x, \lambda_x)/\partial \alpha_x = 0$, we obtain the following generalized eigenvalue problem:

$$\left(K_{xx} + M_{xx}U - M_{xy}V\right)\alpha_x = \lambda_x \left(K_{xx} + \beta I\right)\alpha_x, \tag{23}$$

here we add the term βI with a small positive scalar β to avoid the singularity and is taken as 0.0001 in all the experiments. Therefore, the projection vectors α_x can be computed from the scale-reduced generalized eigenvalue problem in Eq. 23. The projection vectors α_y are obtained by the analogous computational method. Once the projection vector pairs (α_x, α_y) are obtained, the feature selection can be performed in the forms of the low-dimensional transformations using the kernel mapping in Eq. 15. Finally, Eq. 23 can be used for feature selection and classification. After running the LPMDL and KLPMDL algorithms, there existing supervised feature selection learning methods can be effectively and efficiently executed.

3 Experiments and analysis

In this section, we numerically compare the performance of LPMDL and KLPMDL with some existing feature selection methods, i.e. PCA, LPP, NPE and LFDA, for visualization analysis and classification tasks. In LFDA, the affinity matrix is computed by the local scaling method defined in [30]. In the experiments, the RBF kernel $K(x, x^T) = \exp^{\wedge}\left(-||x - x^T||^2/2\sigma^2\right)$ is selected for projecting the points with kernel parameter, $\sigma = 0.01$. The tunable parameters, M_{xx}, M_{yy} and M_{xy}, introduced in the optimization models are respectively set to 1, 1 and 20. We evaluate the learning performance of the proposed algorithm based on several benchmark data sets chose from the UCI ML repository [3] (i.e. *Ionosphere*, *Wisconsin-Breast-Cancer*, *Iris*, *Waveform-5000*, *Vote*, *Heart-statloge*, *Hepatitis*) and the *XOR* data set and the *wood image database* [22]. Chapelle et al. have pointed that each feature selection method performs very well for a particular type of data sets [4], however, it tends to perform poorly for

Table 1 List of the used data sets

Data name	Dimensions	Input samples	# of classes
Ionosphere	34	351	2
Wisconsin-Breast-Cancer	9	699	2
Iris	4	150	3
Waveform-5000	40	5,000	3
Vote	16	435	2
Heart-statlog	13	270	2
Hepatitis	19	155	2
*XOR**	2	20,000	2
*Wood image database**	177	800	2

Data sets indicated by '*' contain within-class multimodal structures

the other varieties of data sets. Thus, the performance of feature selection method is highly associated with the type of data sets and there seems to be no single best method. The wood image set contains intrinsic within-class multimodal structure when they are converted from multi-class problems to the two-class problems by merging some of the classes. The *XOR* data set is also multimodal. For each data set, we choose the parameter κ by *fivefold cross-validation* [14]. Table 1 displays the representations of the data sets used in the experiments. In our experiments, we carry out the experiments on a PC with Intel (R) Pentium (R) D CPU 2.80 GHz 2.79 GHz 512 M. All the used algorithms are implemented in Matlab 7.1.

3.1 Data visualization

We first take the multimodal *XOR* data set used in Sect. 2.1 and the *Iris* and *Waveform-5000* data sets for examples. The *Iris* and *Waveform-5000* contains three types of samples specified by '$*$', ☆ and '$\times$'. We use ('$*$', □) and ('$*$', '$\times$') and ('$\times$', □) to create two-class problems, respectively, and perform dimensionality reduction on them. Results of visualization are shown in the same chart. We choose 1,600 data samples in each class randomly from the *Waveform-5000* data set for experiments. Figure 3 shows the embedded result by each method in the two-dimensional embedding space. For the *XOR* data set, it is obvious that the pair of projection transformations is a further example of the XOR problems. We call this example cross-projections since the data points are obtained by perturbing points originally lying on two sets of nonparallel basis vectors. The projection transformations found by LPMDL can correctly recover the intrinsic multimodal and local structures hided in the given data and achieve a satisfying performance on the used data sets. The embedded results of the original data in different classes can be effectively partitioned from each other in the reduced space found by LPMDL, while LFDA, LPP and NPE tend to mix the projections of the data points. For the *Waveform-5000* data set, the embedding space discovered by NPE, LPP and LFDA are similar to a triangle and more samples data of different classes are mixed with each other, while LPMDL tends to keep in-class data pairs compact and between-class data pairs apart. For the *Iris* data set, the multimodality of the '$\times$'-class can be clearly observed in the results of these methods; however, for the other two classes, NPE, LPP and LFDA mix the embedded samples data. Based on above simulation experiments, LPMDL would be a desirable property in visualization and is found to be more appropriate for embedding the multimodal data than the NPE, LPP and LFDA methods. The experimental results here support the qualitative justification of LPMDL given in Sect. 2.

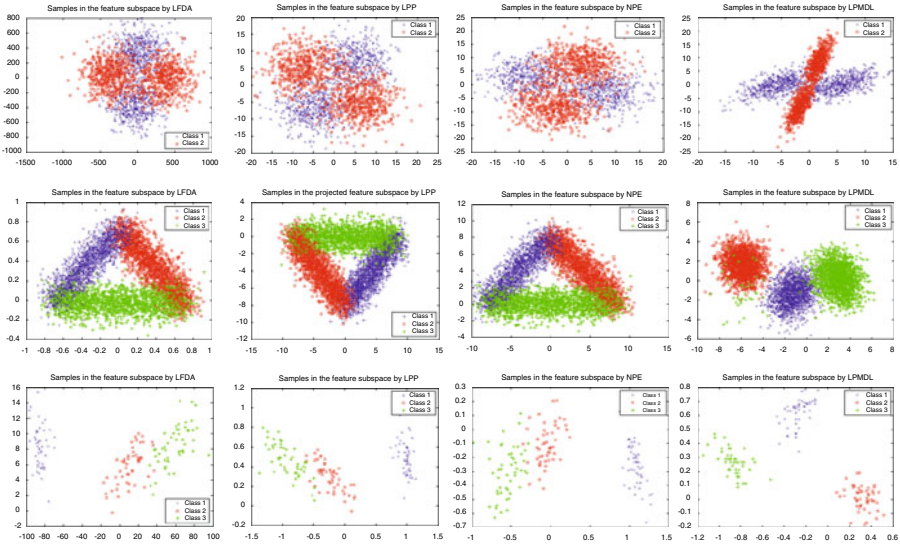


Fig. 3 Results of data visualization. From *top* to *bottom*, the *XOR data set* ($D = 2$, $N = 20000$, $C = 2$, $d = 2$), the *Waveform-5000 data set* ($D = 40$, $N = 5000$, $C = 3$, $d = 4$), the *Iris data set* ($D = 4$, $N = 150$, $C = 3$, $d = 3$). From *left* to *right*, samples in the original spaces and the feature subspaces constructed by LFDA, LPP ($k = 5$), NPE ($k = 5$) and LPMDL ($k = 5$, κ=6, 1.2, 2, respectively). Where, D is the dimensions of the data set, N is the number of instances, C is the number of classes, k is the number of neighbors and d is the selected features

3.2 Experimental results on classification

In this subsection, we investigate the performance of the LPMDL and KLPMDL methods for classification on six benchmark UCI data sets. For avoiding the bias caused by the choice of the classifiers, we introduce the k-nearest-neighbor classifier with Euclidean distance for classification tasks. In short, the classification process has three steps. First, we calculate the image subspace from all the data samples, that is, the points are projected into d-dimensional subspace for each method and create the new input patterns for the experiments; choose the training samples and test samples from the new sample pool, and train a classifier model from the training set; finally, the new sample image is identified and recognized by a k-nearest-neighbor classifier. In the experiments, feature selection is performed by selecting the first d features from the ranking list of features generated by different algorithms, where d is the desired number of selected features specified by users. The performance of LPMDL and KLPMDL are measured by the classification accuracy using the selected features on the testing data. Here we test LPMDL, PCA, LFDA, LPP and NPE on these data sets for comparison. For each data set, we randomly choose the first half of data samples from each class as the training data (*Tr*), and the remaining for testing (*Te*). Before the experiments, we preprocess the data set by adding random vectors chosen from the used data set to ensure that the sample size of *Tr* equals that of *Te* and $m_x = m_y$ for the experiments.

Figure 4 indicates that, in most cases, the performance of LPMDL and KLPMDL is significantly better than those of the PCA, LFDA, LPP and NPE methods as the number of selected features and neighbors increase, especially on the *XOR*, *Ionosphere*, *Hepatitis* and *Heart-statlog* data sets. Moreover, LPMDL and KLPMDL tend to remain stable for a wide range of reduced dimensions. From Figs. 4a, b, d and f, LPMDL and KLPMDL almost always

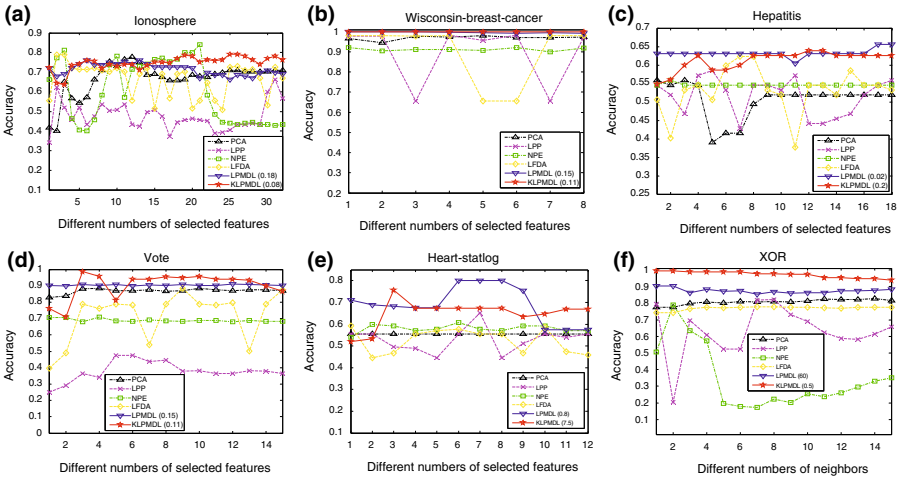


Fig. 4 Accuracy versus different numbers of selected features or neighbors on the six UCI data sets: **a** on *Ionosphere*; **b** on *Wisconsin-Breast-Cancer*; **c** on *Hepatitis*; **d** on*Vote*; **e** on *Heart-statlog*; **f** on *XOR*

achieve the highest classification accuracy, which is comparable to that of PCA. As shown in Figs. 4c and e, our method performs better than the other methods significantly. For the *Ionosphere* data set, with increasing of the selected features, the classification accuracies of PCA and LPP are comparative, and both superior to those of NPE and LFDA. As described in Figs. 4b and d, LPP and LFDA work better than NPE, especially on the *Vote* data set. For the *XOR* data set, PCA and LFDA perform better than LPP and NPE methods in a great extent, but both inferior to LPMDL and KLPMDL. What is even more, LPP and NPE cannot keep stable for a wide range of selected features and are sensitive to the selected features. LFDA works poorly on the *Heart-statlog* data set, and the performance is almost always inferior to the other methods. The values in the brackets are the parameter κ here and in later sections.

Moreover, we show the averaged accuracy obtained by PCA, LPP, NPE, LFDA, LPMDL and KLPMDL, over 20 random splits of training samples, in Table 2. The classification accuracy in Fig. 4 and Table 2 indicate that when the number of selected features is smaller, both LPMDL and KLPMDL can achieve the comparative performance to the classical PCA, LPP and LFDA methods. On the contrary, PCA, LPP, LPMDL and KLPMDL are more robust to the number of the selected features.

To investigate the runtime performance of these algorithms under different numbers of selected features and nearest neighbors, we perform PCA, LPP, NPE, LFDA and LPMDL on the *Hepatitis*, *XOR* and *Heart-statlog* data sets and the experimental results are plotted in Fig. 5, which represents that the runtime performance of LPMDL is generally superior to that of NPE and LPMDL works slightly slower than the LPP, PCA and LFDA methods. For the *XOR* data set, NPE need much computation time than the other methods and increase faster with the increasing of the number of nearest neighbors.

3.3 Recognition results on wood image database

In this study, we will investigate the proposed algorithm for wood defects recognition. In the experiments, local binary pattern (LBP) [18] is used to select the texture features from the wood image set, including 400 negative samples labeled by -1 and 400 positive ones

Table 2 The averaged accuracy of the several algorithms on the selected data sets and the methods are listed in rows

Data name	PCA	LPP	NPE	LFDA	LPMDL	KLPMDL
Ionosphere	0.6687 ($k = 5$)	0.4691 ($k = 5$)	0.5562 ($k = 5$)	0.6237 ($k = 5$)	0.7085 ($k = 5, \kappa = 0.18$)	0.7393 ($k = 5, \kappa = 0.08$)
Wisconsin-Breast-Cancer	0.9699 ($k = 5$)	0.8954 ($k = 5$)	0.9118 ($k = 5$)	0.8986 ($k = 5$)	0.9961 ($k = 5, \kappa = 0.15$)	0.9983 ($k = 5, \kappa = 0.11$)
Hepatitis	0.5065 ($k = 5$)	0.4805 ($k = 5$)	0.5455 ($k = 5$)	0.5447 ($k = 5$)	0.6419 ($k = 5, \kappa = 0.02$)	0.6356 ($k = 5, \kappa = 0.2$)
Vote	0.8694 ($k = 5$)	0.3794 ($k = 5$)	0.6895 ($k = 5$)	0.7167 ($k = 5$)	0.9044 ($k = 5, \kappa = 0.15$)	0.9032 ($k = 5, \kappa = 0.11$)
Heart-statlog	0.5556 ($k = 5$)	0.5326 ($k = 5$)	0.5764 ($k = 5$)	0.4968 ($k = 5$)	0.7257 ($k = 5, \kappa = 0.8$)	0.6894 ($k = 5, \kappa = 7.5$)
XOR	0.7760 ($k = 5$)	0.6315 ($k = 5$)	0.6475 ($k = 5$)	0.7190 ($k = 5$)	0.8770 ($k = 5, \kappa = 60$)	0.9508 (k=5, κ=0.5)

The numbers in the bracket are respectively the numbers of neighbors and the values of tunable parameters κ trained by *fivefold cross-validation*

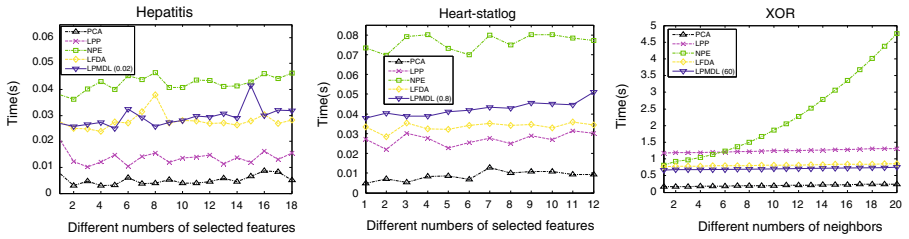


Fig. 5 Runtime performance versus different numbers of selected features or neighbors on the *Hepatitis*, *Heart-statlog* and *XOR* data sets

labeled by 1, chose from the wood database [22], which consists of a large number of pine boards with ground truth classifications for each defect. The results reported here are based on a set of 438 samples images with over 200 labeled defects. The imaging resolution has been 0.5 mm, and a color line-scan camera has been used for image acquisition. In the experiments, we randomly choose first half of data samples for training and the remaining as the test data. In both training and test sets, each class has 200 samples randomly selected from the image set. We repeat this process 20 times and compute the average accuracy. Different pattern classification technologies have been applied for wood defects recognition based on the real-world image data base, e.g. [21]. Here, we apply our proposed method and 5-nearest-neighbor classifier (5-NN) for defects recognition on the wood feature set. Here, we first apply the existing PCA, LPP, NPE, LFDA and our LPMDL and KLPMDL methods to the feature set for 2D visualization and then evaluate the classification performance of the proposed method in the real-world data set further.

Figure 6 shows the results of the image data visualization. Analogous to Sect. 3.1, we obtain a satisfying result and the cross-transformations learned by LPMDL can correctly recover the intrinsic structure information of the wood feature set and learn a optimal embedding space, under which the positive and negative samples can be effectively partitioned from each other. Figure 7 shows the plots of the recognition accuracy and the runtime performance vs. different number of selected features on the feature set. Table 3 gives the averaged

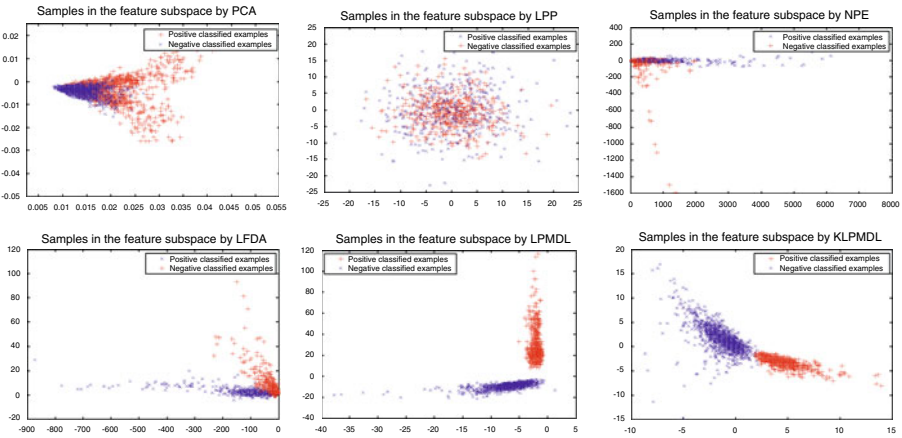


Fig. 6 2D Data visualization of the *wood image database*

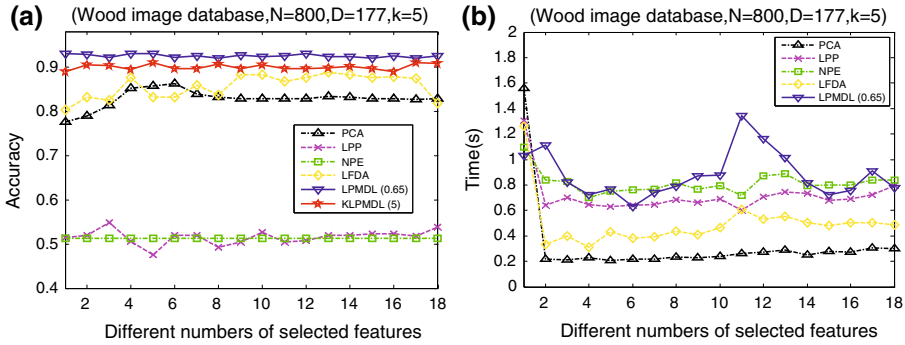


Fig. 7 Recognition accuracy and runtime performance versus different numbers of selected features on the *wood image database*

Table 3 The averaged accuracy of the several algorithms under different number of selected features on the *wood image database*

Data name	PCA	LPP	NPE	LFDA	LPMDL	KLPMDL
Wood image database	0.8285 ($k = 5$)	0.5155 ($k = 5$)	0.5129 ($k = 5$)	0.8565 ($k = 5$)	0.9248 ($k = 5, \kappa = 0.65$)	0.8997 ($k = 5, \kappa = 5$)

accuracy under different numbers of selected features on the feature set. As shown in Fig. 7a and Table 3, LPMDL and KLPMDL work well on the data set and perform the comparative results to PCA and LFDA, which again verifies the usefulness of the proposed methods for pattern classification. Relatively, LPP and NPE perform poorly on the data set. Figure 7b displays the running time (here, we compute the time in seconds) of these methods with the increasing of the selected features, from which we can find the runtime performance of our LPMDL method is close to those of LPP and NPE. PCA and LFDA work fast on the data set.

4 Conclusions and outlooks

In this paper, we focused on the supervised feature selection problem where samples were accompanied with the class information and proposed a novel Locality preserving multimodal discriminative learning algorithm, namely LPMDL, for feature selection. LPMDL can optimally embed the labeled multimodal data appropriately and can capture the local neighborhood information of the data manifold in a certain sense, that is, if two points are mutually local neighbors in the original space, then the neighborhood relationship still holds in the reduced feature space. LPMDL has an analytical form of the embedding transformation, which can be effectively and easily computed based on eigen-decomposition. By defining the new formulations, our approach is interesting and has some distinctive advantages over some existing feature selection techniques.

We test PCA, NPE, LPP, LFDA and our method in data visualization and classification experiments based on some benchmark UCI data sets, the XOR data set and the real-world wood image database. For visualization, LPMDL can separate the projections of data in different classes from each other in addition to preserving the local and multimodal structures due to the optimal embedding space that the samples data are projected on. The test results show that PCA, NPE, LPP, LFDA tend to overlap the projections of the data in different classes if they are close in the original space. For classification, LPMDL can select the good features from the original set and train a high-performance nearest-neighbor classifier model. The classification accuracy show LPMDL achieves the comparable or even better learning performance to some classical methods. Moreover, the runtime performance of these methods is comparative when the dimensionality of the data is not very high.

This is our current research topic. It is interesting to investigate whether we can improve the performance by introducing the unlabeled samples for representation and extending LPMDL to the semi-supervised learning case. Furthermore, the learning performance of the kernelized LPMDL heavily depends on the choice of the kernels and its parameters. Thus, how to optimally determine the kernels and estimate the kernel parameters for the nonlinear learning method needs to be explored in the future work.

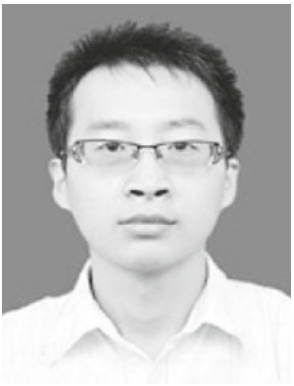
Acknowledgments This research was supported by the National Natural Science Foundation of China under Grant No. 30671639, the Natural Science Foundation of Jiangsu province of China under Grant No. BK2009393, the Innovation Program Foundation of Jiangsu Province of China under Grant No. 164070265, the Innovation Program Foundation of Nanjing Forestry University under Grant No. 2009106 and the Scientific Research Foundation of Jiangsu Province of China under Grant No. CX09S_013Z.

References

1. Belkin M, Niyogi P (2001) Laplacian eigenmaps and spectral techniques for embedding and clustering. Advances in neural information processing system. MIT Press, Cambridge 585–591
2. Belkin M, Niyogi P (2003) Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput 15(6): 1373–1396
3. Blake C, Keogh E, Merz CJ (1998) UCI repository of machine learning databases. Department of Information and Computer Science, University of California, Irvine. http://www.ics.uci.edu/~mlearn/MLRepository.html
4. Chapelle O, Schölkopf B, Zien A (eds) (2006) Semi-supervised learning. MIT Press, Cambridge
5. Chung FRK (1997) Spectral graph theory. AMS, pp 43–107
6. Dy JGC, Brodley E (2004) Feature selection for unsupervised learning. J Mach Learn Res 5(August): 845–889
7. Dy JG, Brodley CE, Kak AC, Broderick LS, Aisen AM (2003) Unsupervised feature selection applied to content-based retrieval of lung images. IEEE Trans Pattern Anal Mach Intell 25(3): 373–378

8. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3(March): 1157–1182
9. Hastie T, Tibshirani R, Friedman J (eds) (2001) The elements of statistical learning: data mining, inference, and prediction. Springer, New York pp 534–553
10. He X, Niyogi P (2003) Locality preserving projections. Advances in neural information processing systems. MIT Press, Cambridge 585–591
11. He X, Deng C, Yan SC, Zhang HJ (2005) Neighborhood preserving embedding. In: Proceeding of international conference on computer vision. IEEE CS Press, Washington, DC, pp 1208–1213
12. He X, Yan S, Hu Y, Niyogi P, Zhang H (2005) Face recognition using Laplacianfaces. IEEE Trans Pattern Anal Mach Intell 27(3): 328–340
13. Kim TK, Kittler J (2005) Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image. IEEE Trans Pattern Anal Mach Intell 27(3): 318–327
14. Lin YY, Liu TL, Chen HT (2005) Semantic manifold learning for image retrieval. In: Proceedings of the ACM conference on multimedia. ACM Press, Singapore, pp 249–258
15. Mika S, Ratsch G, Weston J et al (1999) Fisher discriminant analysis with kernels. In: Hu YH, Larsen J, Wilson E, Douglas S (eds) Proceeding of the IEEE international workshop on neural networks for signal processing. IEEE Press, Madison, pp 41–48
16. Min W, Lu K, He X (2004) Locality pursuit embedding. Pattern Recognit 37(4): 781–788
17. Mokbel MF, Aref WG, Grama A (2003) Spectral LPM: an optimal locality-preserving mapping using the spectral (not fractal) order. In: Proceedings of 19th international conference on data engineering. IEEE Press, Bangalore, India, pp 699–701
18. Ojala T, Pietikäinen M, Mäenpää T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 24(7): 971–987
19. Saul LK, Roweis ST (2003) Think globally, fit locally: unsupervised learning of low dimensional manifolds. J Mach Learn Res 4(December): 119–155
20. SchÄolkopf B, Smola AJ (2002) Learning with kernels. MIT Press, Cambridge 25–55
21. Silven O, Niskanen M, Kauppinen H (2003) Wood inspection with non-supervised clustering. Mach Vis Appl 13(5–6): 275–285
22. Sommardahl O, Usenius A (1999) Wood samples image database. VTT Building Technology. http://www.ee.oulu.fi/~olli/Projects/Lumber.Grading.html
23. Song GJ, Cui B, Zheng BH, Xie Kq, Yang DQ (2008) Accelerating sequence searching: dimensionality reduction method. Knowl Inf Syst 20(3): 301–322
24. Sugiyama M (2007) Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. J Mach Learn Res 8(May): 1027–1061
25. Tsang IW, Kwok JT (2003) Distance metric learning with kernels. In: Proceeding of international conference on artificial neural networks, pp 126–129
26. Vlachos M, Domeniconi C, Gunopulos D (2002) Non-linear dimensionality reduction techniques for classification and visualization. In: Proceedings of international conference on knowledge discovery and data mining. ACM Press, Edmonton, Canada, pp 645–651
27. Verbeek JJ, Roweis ST, Vlassis N (2003) Non-linear CCA and PCA by alignment of local models. Advances in neural information processing systems. MIT Press, Cambridge 297–304
28. Xiang SM, Nie FP, Zhang CS, Zhang CX (2006) Spline embedding for nonlinear dimensionality reduction. In: Proceedings of European conference on machine learning. Lecture Notes in Computer Science, Berlin, Germany, pp 825–832
29. Xiang SM, Nie FP, Song YQ, Zhang CS, Zhang CX (2009) Embedding new data points for manifold learning via coordinate propagation. Knowl Inf Syst 19(2): 159–184
30. Zelnik-Manor L, Perona P (2005) Self-tuning spectral clustering. In: Saul LK, Weiss Y, Bottou L (eds) Advances in neural information processing systems. MIT Press, Cambridge pp 1601–1608
31. Zhang DQ, Chen SC (2003) Clustering incomplete data using kernel-based fuzzy c-means algorithm. Neural Process Lett 18(3): 155–162
32. Zhao HT, Sun SY, Jing ZL, Yang JY (2006) Local structure based supervised feature extraction. Pattern Recognit 39(8): 1546–1550
33. Zhou D, Bousquet O, Lal T, Weston J, Schölkopf B (2004) Learning with local and global consistency. Advances in neural information processing systems. MIT Press, Cambridge, 321–328

Author Biographies



Zhao Zhang has been a master candidate in Department of Computer Science and Technology, Nanjing Forestry University, China. His current interests include machine learning and pattern recognition. He has written more than ten research papers related to his interests, of which several have been accepted in journals and international conferences and one of them has been selected as the excellent paper in China National Computer Conference, 2009. His research was supported by the Natural Science Foundation of Jiangsu province of China under Grant No. BK2009393, the Scientific Research Foundation of Jiangsu Province of China under Grant No.CX09S_013Z, the Scientific Innovation Foundation of Nanjing Forestry University under Grant No. 2009106 and the Innovation Program Foundation of Jiangsu Province of China under Grant No. 164070265.

Ning Ye has received his Ph.D degree from the Department of Computer Science, Southeast University, China. He is currently a Professor in the Department of Computer Science, Nanjing Forestry University, China. His current research focus is on bioinformatics, machine learning, data mining and pattern recognition. His research has strong direct links to industrial user and have attracted many funding, which include the National Science Foundation of China (No. 302071048, No. 30671639, No. 5BZX025), and the Natural Science Foundation of Jiangsu province of China (No. BK2005134, No. BK2009393). One of his research papers about the Fast Algorithms on SVM bas been selected as the one and only one excellent paper in the 1st China conference on classification, 2004. At present, he is a member of Artificial Intelligence Association in Jiangsu province of China and a senor member of China Computer Federation.