Engineering Applications of Artificial Intelligence 75 (2018) 94–101

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

A global manifold margin learning method for data feature extraction and classification

Bo Li [a,b,*], Wei Guo [a,b], Xiao-Long Zhang [a,b]

[a] School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, 430065, PR China
[b] Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan, 430065, PR China

ARTICLE INFO

Keywords:
Feature extraction
Supervised manifold learning
Manifold margin

ABSTRACT

This paper presents a global manifold margin learning approach for data feature extraction or dimensionality reduction, which is named locally linear representation manifold margin (LLRMM). Provided that points locating on one manifold are of the same class and those residing on the corresponding manifolds are varied labeled, LLRMM is desired to identify different manifolds, respectively. In the proposed LLRMM, it firstly constructs both a between-manifold graph and a within-manifold graph. In the between-manifold graph, for any point, its k nearest neighbors and itself must belong to different manifolds. However, any node and its neighborhood points should be on the same manifold in the within-manifold graph. Then we use the minimum locally linear representation trick to reconstruct any node with their corresponding k nearest neighbors in both graphs, from which a between-manifold graph scatter and a within-manifold graph scatter can be reasoned, followed by a novel global model of manifold margin. At last, a projection will be explored to map the original data into a low dimensional subspace with the maximum manifold margin. Experiments on some widely used face data sets including AR, CMU PIE, Yale, YaleB and LFW have been carried out, where the performance of the proposed LLRMM outperforms those of some other methods such as kernel principal component analysis (KPCA), non-parametric discriminant analysis (NDA), reconstructive discriminant analysis (RDA), discriminant multiple manifold learning (DMML) and large margin nearest neighbor (LMNN).

1. Introduction

For image pattern classification besides face recognition, it often rewrites the original data to high dimensional vectors. For instance, an appearance-based face image with size 80*80 can be transformed to a 6400-dimension vector. So it is required to extract discriminant features from the high dimensional vectors before making classification to them, which will contribute to improving recognition performance with low computational expense. Currently, researchers have reported many dimensionality reduction or feature extraction methods where both the linear and the nonlinear models are all involved (Wang et al., 2016; Yu et al., 2016; Sadatnejad and Ghidary, 2016; Motta et al., 2015). Moreover, they have been widely used in many applications with convincing performances (Zhang et al., 2016b, a; Sun et al., 2013; Zhang et al., 2016c).

As a traditional linear feature extraction method, principal component analysis (PCA) aims to locate a subspace where the covariance of all the original data can be maximized (Jolliffe, 2002). Meanwhile, it should be noted that no class information is taken into account in PCA, thus the supervised information do not play its role in the following feature extraction and classification (Yang and Zhang, 2008). However, another classical linear method, i.e. linear discriminant analysis (LDA), constructs an objective function by taking full consideration of data class labels (Kim et al., 2011; Martinez and Kak, 2001). In general, LDA projects the original data into a subspace with the maximum between-class apartness and the minimum within-class compactness.

The above mentioned methods concentrate on global linear structure of the original data and fail to dig nonlinear information hidden in them, which has been validated to be useful for dimensionality reduction (Tenenbaum et al., 2000; Roweis and Saul, 2000). Thus some nonlinear learning techniques are prevailing. As one kind of famous nonlinear models, neural networks have been attracting more and more attentions. In 1996, Huang (2004a, b) systematically concluded the theory of neural networks and their applications to pattern recognition. Moreover, neural networks have also been used to find polynomial roots (Huang, 2004a, b; Huang et al., 2005). On the basis of the traditional neural networks, radial basis probabilistic neural network (RBPNN) is constructed by a constructive hybrid strategy (Huang and Du, 2008). The

* Corresponding author.
E-mail address: liberol@126.com (B. Li).

https://doi.org/10.1016/j.engappai.2018.08.004
Received 25 September 2017; Received in revised form 21 March 2018; Accepted 6 August 2018

proposed RBPNN has also been introduced for biometric identification (Shang et al., 2006; Zhao et al., 2004). However, neural networks optimize weights between nodes with so many iterations that much more computational cost will be paid when dealing with real world data.

In addition, kernel transformations, i.e. kernel principal component analysis (KPCA) (Scholkopf et al., 1998; Wen et al., 2012) and kernel Fisher discriminant analysis (KFDA) (Yang et al., 2004), are also presented to implicitly map observations into a space with high dimensionality, where they can be linearly classified. However, not local geometry but global structure information is approached from high-dimensional data using both KPCA and KFDA. Under such circumstances, manifold learning is put forward to explore manifold geometry hidden in the high dimensional data by locality learning.

It is well known that many manifold learning methods have been presented during last decade (Tenenbaum et al., 2000; Roweis and Saul, 2000; Saul and Roweis, 2003; Donoho and Grams, 2003; Belkin and Niyogi, 2003; Zhang and Zha, 2005; Weinberger and Saul, 2006; Lin and Zha, 2008). Among them, all the local patches on manifold are determined using k nearest neighbors or super-ball criterion, where linear tricks can be well performed to find the locality of manifold. Moreover, by taking advantage of data class information, some modifications are also made to them. For example, marginal Fisher analysis (MFA) (Xu et al., 2007; Yan et al., 2007) and discriminant multi-manifold learning (DMML) (Lu et al., 2013) construct two different graphs to represent the within-class compactness and the between-class separability, respectively. Additionally, by carrying out traditional LDA to all the local patches, local Fisher discriminant analysis (LFDA) (Sugiyama, 2006), non-parametric discriminant analysis (NDA) (Li et al., 2009) and reconstructive discriminant analysis (RDA) (Yang et al., 2008; Chen and Jin, 2012) maximize the trace ratio of the local inter-class graph scatter to the local intra-class graph scatter to find an optimal subspace. All the objective functions of these methods are under framework of trace ratio, which often incurs small sample size problem (Kim et al., 2011). To prevent the problem, locality sensitive discriminant analysis (LSDA) defines a local margin, which can be deduced to difference between the local inter-class graph scatter to the local intra-class graph scatter (Cai et al., 2007). However, it just pays more attentions to separateness between local patches. Another manifold learning based dimensionality reduction method titled local discriminant embedding (LDE) models two graphs based on data neighborhood and class relation, and then two graph Laplacians are used to find low-dimensional embeddings (Chen et al., 2005). Recently, a novel manifold method, i.e. t-distributed stochastic neighbor embedding (t-SNE), has also been focused on and high performances have been achieved by using it for feature extraction and pattern recognition (van der Maaten, 2014). However, both LDE and t-SNE show no concern on global manifold margin, which can characterize the total apartness of all the manifolds.

Thus how to globally measure all the manifolds' margin still needs further demonstration. In this paper, we will propose a globally defined manifold margin metric with locally linear representation strategy that can be introduced to measure apartness among all the manifolds. Based on the proposed manifold margin, a locally linear representation manifold margin (LLRMM) method will be put forward for multi-manifold identification. The main contributions of the proposed LLRMM are listed below:

(1) Although manifold margin is firstly proposed in the conference paper with a simply version (Li et al., 2015), in this paper, it is also presented to characterize the global apartness among different manifolds with a graphical illustration. Moreover, more details are offered either from the construction of three kind graphs and their corresponding scatters using the minimum linear representation trick or from theoretical derivations of the proposed manifold margin.

(2) Based on the proposed manifold margin, LLRMM is also put forward to extract discriminant features accompanying with its outline.

(3) Much more experiments on benchmark face data such as AR, CMU PIE, Yale, YaleB and LFW are carried out to obtain the statistics results including mean recognition rates and standard deviations, from which the performance of the proposed LLRMM can be validated.

The rest of the paper is organized as follows. The proposed algorithm is described and justified in Section 2. Section 3 presents the experimental results on face data accompanying with some discussions. At last, it draws some conclusions and makes some expectations for the future work in Section 4.

2. Locally linear representation manifold margin

2.1. Motivation

Recently, many supervised extensions have been made to LLE to deal with data classification problem. In these methods, some are proposed by combining LDA to LLE (Zhang et al., 2004, 2006; Pang et al., 2006; de Ridder et al., 2004; Li et al., 2008), some other take class information into account to direct the construction of local graph. However, when constructing k nearest neighbor graph, some distances between varied labeled nodes may be shorter than those between points sampled from the same class, which will lead to wrong neighborhood selection for discriminant analysis. In order to overcome the problem, some methods are put forward either by adjusting distances between nodes or by just selecting neighbors from nodes with the same class (de Ridder et al., 2003; Wen and Jiang, 2006; Zhang and Zhao, 2007; Zhao and Zhang, 2009; Hui and Wang, 2008; Zhao et al., 2005; Han et al., 2008). Assume that points locating on one manifold are of the same class and those residing on the corresponding manifolds are sampled from varied labeled data, these approaches are aiming to construct a k nearest neighbor graph to characterize the within-manifold data. However, they ignore to set up another k nearest neighbor graph which is composed of the between-manifold data. Thus both the within-manifold graph and the between-manifold graph will be constructed, by which a manifold margin metric can be globally proposed to quantify the apartness among different manifolds. At the same time, following the within-manifold graph and the between-manifold graph, a total-manifold graph will also be introduced to measure locality of all the samples without considering manifold label information. Compared to the original LLE (Lawrence, 2001), which is an unsupervised dimensionality reduction method, the proposed LLRMM takes manifold label information into account to construct a between-manifold graph and a within-manifold graph, respectively. In the between-manifold graph, any node and its k nearest neighbors must belonging to different manifolds. Thus the distances between multiple manifolds may exist in the between-manifold graph, which shows close relation to the expected global manifold margin.

Fig. 1 illustrates the proposed LLRMM method, where binary classification problem is involved. In Fig. 1, there are two differently labeled manifolds M1 and M2. For one point in a manifold M1, its four within-manifold nearest neighbors are selected to be composed of its within-manifold graph. Meanwhile, it also chooses other four nearest neighbors on another manifold to consist of its between-manifold graph. However, from the left sub-figure in Fig. 1, it can be found that two manifold data are mixed together and cannot be distinguished in the original high dimensional space. So in order to identify these two manifold data, it is expected to find a low dimensional subspace to maximize the manifold margin shown in right subfigure in Fig. 1.

But how to define the global manifold margin is still a problem. In the following, it will be reasoned from the within-manifold graph scatter and the between-manifold graph scatter using the minimum linear representation technique.

2.2. Locally linear representation weights

When constructing the within-manifold graph, both local geometry and manifold label information are all employed. On one hand, any node and its neighborhood should be on a manifold in the within-manifold graph. On the other hand, its local neighborhood should be composed

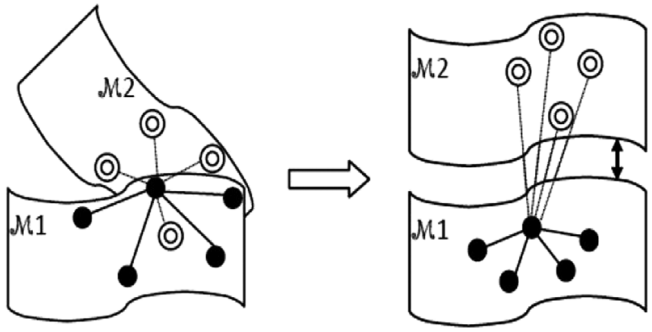

Fig. 1. Illustration of the proposed LLRMM.

of its k nearest neighbors. For any node in the within-manifold graph, it can be represented by its local neighborhood points. Moreover, it often uses the minimum representation error to obtain the optimal weights between them. The objective function is described as follows.

$$\varepsilon\left((w_w)_i\right) = \min\left\|X_i - \sum_{j=1}^{k}(W_w)_{ij}X_j\right\|^2 \tag{1}$$

where $X_j(j = 1,2,\ldots,k)$ are the k within-manifold neighborhood neighbors of X_i. Because the locally linear representation weight $(W_w)_{ij}$ is under sum-to-one constraint, i.e. $\sum_{j=1}^{k}(W_w)_{ij} = 1$, Eq. (1) can be rewritten to the following form.

$$\begin{aligned}\varepsilon\left((W_w)_i\right) &= \min\left\|\sum_{j=1}^{k}(W_w)_{ij}X_i - \sum_{j=1}^{k}(W_w)_{ij}X_j\right\|^2 \\ &= \min\left\|\sum_{j=1}^{k}(W_w)_{ij}(X_i - X_j)\right\|^2\end{aligned} \tag{2}$$

So it can also be expressed with a local gram matrix.

$$\begin{aligned}\varepsilon\left((W_w)_i\right) &= \min\left\{\sum_{j=1}^{k}(W_w)_{ij}(X_i - X_j)\cdot\sum_{t=1}^{k}(W_w)_{it}(X_i - X_t)\right\} \\ &= \min\sum_{j=1}^{k}\sum_{t=1}^{k}(W_w)_{ij}(W_w)_{it}G_{jt}\end{aligned} \tag{3}$$

where $G_{jt} = \left(X_i - X_j\right)\cdot\left(X_i - X_t\right)$ denotes the local gram matrix.

Using Lagrange multiplier, the locally linear representation weights in the within-manifold graph can be obtained.

Firstly, a Lagrange function can be formed as:

$$L = \sum_{j=1}^{k}\sum_{t=1}^{k}(W_w)_{ij}(W_w)_{it}G_{jt} - \lambda(\sum_{j=1}^{k}(W_w)_{ij} - 1)$$

Then let

$$\frac{\partial L}{\partial W_w} = 0$$

Thus, the locally linear representation weights in the within-manifold graph can be deduced to the following formulation.

$$(W_w)_i = \frac{\sum_{t=1}^{k}G_{jt}^{-1}}{\sum_{m=1}^{k}\sum_{l=1}^{k}G_{lm}^{-1}} \tag{4}$$

Based on Eq. (4), we can also define the locally linear representation weights between all the nodes in the within-manifold graph as Eq. (5).

$$(W_w)_i = \begin{cases}\dfrac{\sum_{t=1}^{k}G_{jt}^{-1}}{\sum_{m=1}^{t}\sum_{l=1}^{k}G_{lm}^{-1}} & X_j \in WithinN(X_i) \\ 0 & otherwise\end{cases} \tag{5}$$

where $WithinN(X_i)$ contains all the neighborhood points of node X_i in the within-manifold graph. Thus for all the nodes in the within-manifold graph, their corresponding locally linear representation weights can be found, which consist of within-manifold weight matrix W_w.

For each node in the between-manifold graph, the similar process is repeated except that its neighborhood points should be selected from different manifolds. Then we can achieve a locally linear representation weight matrix W_b in the between-manifold graph, where each row is the corresponding $(W_b)_i, i = 1,2,\ldots,n$.

$$(W_b)_i = \begin{cases}\dfrac{\sum_{t=1}^{k}G_{jt}^{-1}}{\sum_{m=1}^{t}\sum_{l=1}^{k}G_{lm}^{-1}} & X_j \in BetweenN(X_i) \\ 0 & otherwise\end{cases} \tag{6}$$

where $BetweenN(X_i)$ signs the between-manifold neighborhood of node X_i and W_b is the weight matrix in the between-manifold graph with row $(W_b)_i, i = 1,2,\ldots,n$.

In addition, when constructing the total-manifold graph, just those with k shortest Euclidean distances to node X_i are taken as its total-manifold neighborhood and the locally linear representation weights are listed in the following.

$$(W_t)_i = \begin{cases}\dfrac{\sum_{t=1}^{k}G_{jt}^{-1}}{\sum_{m=1}^{t}\sum_{l=1}^{k}G_{lm}^{-1}} & X_j \in TotalN(X_i) \\ 0 & otherwise\end{cases} \tag{7}$$

where $TotalN(X_i)$ is the total-manifold neighborhood of node X_i and W_t is the total-manifold graph weight matrix, which is composed of $(W_t)_i, i = 1,2,\ldots,n$.

2.3. Manifold Margin

In the above Subsection, three graphs including the within-manifold graph, the between-manifold graph and the total-manifold graph are all quantified with their locally linear representation weights between nodes, which can be used to character the within-manifold data, the between-manifold data and all the data. Moreover, based on these three graphs, a within-manifold graph scatter, a between-manifold graph scatter and a total-manifold graph scatter can be defined, respectively.

In Eq. (1), we have applied the least locally linear representation trick to optimize the weights between a node and its within-manifold graph neighbors. Using the quadratic form, Eq. (1) can also be rewritten to Eq. (8).

$$\begin{aligned}&\varepsilon\left((W_w)_i\right) = \min\left\|\sum_{j=1}^{k}(W_w)_{ij}X_i - \sum_{j=1}^{k}(W_w)_{ij}X_j\right\|^2 \\ &= \min tr\{(X_i - \sum_{j=1}^{k}(W_w)_{ij}X_j)(X_i - \sum_{j=1}^{k}(W_w)_{ij}X_j)^T\} \\ &= \min tr\{\sum_{i,j}(U_w)_{ij}\left(X_i \bullet X_j\right)\}\end{aligned} \tag{8}$$

where $U_w = (I - W_w)(I - W_w)^T$.

Moreover, for all the points, it can be changed into another formulation with matrix form.

$$\varepsilon\left(W_w\right) = \min tr(XU_wX^T) \tag{9}$$

Eq. (9) aims to find the minimum representation error among the within-manifold data by using a constrained optimized objective function, which can also be viewed to the compactness of the within-manifold data. Thus it is introduced to characterize the within-manifold graph scatter, which is defined below.

$$S_w = XU_wX^T \tag{10}$$

Similar to the within-manifold graph scatter, the between-manifold graph scatter S_b and the total-manifold graph scatter S_t will be formulated, respectively.

$$S_b = XU_bX^T \tag{11}$$

where $U_b = (I - W_b)(I - W_b)^T$

$$S_t = XU_tX^T \tag{12}$$

where $U_t = (I - W_t)(I - W_t)^T$

In the following, a novel manifold margin S_M will be globally modeled to measure the apartness among different manifolds.

$$S_M = \sum_{i,j} d(M_i, M_j) - \sum_i tr(M_i) \tag{13}$$

where $M_i$marks the ith manifold and $d(M_i, M_j)$ represents distance between manifold M_i to manifold M_j, $tr(M_i)$ is taken to represent the scale of manifold M_i.

Based on Eq. (13), it will find that the total minimum distances between all the points on manifold M_i to other manifolds $M_j(j = 1, 2, \ldots, c, i \neq j)$ can be taken as the expected manifold distances $d(M_i, M_j)$. Thus, we have:

$$d(M_i, M_j) = \sum_{i,j} \{\min d(X_i, M_j)\} \tag{14}$$

Moreover, it should be pointed out that the shortest distance between point X_i and other manifolds $M_j(j = 1, 2, \ldots, c, i \neq j)$ can also be replaced with the distance between point X_i and the weighted mean of its k between-manifold neighborhood points. Thus Eq. (14) can be rewritten to:

$$d(M_i, M_j) = \sum_i \left\| X_i - \sum_{j=1}^{k} (W_b)_{ij} X_j \right\|^2 \tag{15}$$

where $X_j(j = 1, 2, \ldots, k)$ are k between-manifold neighborhood points for point X_i.

Therefore, the proposed manifold margin has the following form.

$$S_M = \sum_i \left\| X_i - \sum_{j=1}^{k} (W_b)_{ij} X_j \right\|^2 - \sum_i tr(M_i) \tag{16}$$

In the between-manifold graph, $\sum_i \left\| X_i - \sum_{j=1}^{k} (W_b)_{ij} X_j \right\|^2$ is just the between-manifold graph scatter, thus Eq. (16) can be expressed to:

$$S_M = S_b - \sum_i tr(M_i) \tag{17}$$

In addition, Eq. (1) displays that a point can be represented with its within-manifold neighborhood points, where the minimum representation error $\epsilon\left((W_w)_i\right)$ demonstrates the within-manifold data compactness. Thus the total minimum representation errors in the within-manifold graph can be used to weigh its scale. So Eq. (18) will be obtained.

$$\sum_i tr(M_i) = \sum_i \epsilon\left((W_w)_i\right) \tag{18}$$

With matrix form, it should be:

$$\sum_i tr(M_i) = tr(XU_wX^T) \tag{19}$$

Thus the compactness of all manifolds can be approached by the within-manifold graph scatter.

At last, the predefined manifold margin in the high dimensional space will be stated to the following form:

$$S_M = S_b - S_w = X(U_b - U_w)X^T \tag{20}$$

2.4. Justification

For multiple manifolds identification, it should be explored a subspace where different labeled manifold data will be easily separable. It will come true by maximizing the manifold margin in the low dimensional space, i.e.

$$\max\{Y(U_b - U_w)Y^T\} \tag{21}$$

However, instead of the within-manifold graph scatter, the total-manifold graph scatter, which can be introduced to represent the inherent geometry of all the manifolds, will be minimized in the desired subspace. Thus, all manifold locality will be preserved as the traditional manifold leaning methods.

$$\min\{YU_tY^T\} \tag{22}$$

Based on the above analysis, the low dimensional embeddings will be obtained by satisfying two objective functions. It can be concluded that they should be the solutions of the following multi-objective optimization problem:

$$\begin{cases} \max tr\left\{Y(U_b - U_w)Y^T\right\} \\ \min tr\left\{YU_tY^T\right\} \end{cases} \tag{23}$$

To solve Eq. (23), we first change it into a single objective function as follows:

$$\max \frac{tr\left\{Y(U_b - U_w)Y^T\right\}}{tr\left\{YU_tY^T\right\}} \tag{24}$$

It must be noted that traditional manifold learning methods often encounter out-of-sample problem, thus a linear transformation between the original data and their embeddings is recommended, i.e. $Y = A^TX$. Moreover, the orthogonality is also constrained to the linear transformation, i.e. $A^TA = I$, so Eq. (24) can be represented to the following Fisher criterion under orthogonal constraint:

$$\max \frac{tr\left\{A^TX(U_b - U_w)X^TA\right\}}{tr\left\{A^TXU_tX^TA\right\}} = \max \frac{tr(A^TS_MA)}{tr(A^TS_tA)} \tag{25}$$

$$s.t. \quad A^TA = I$$

Eq. (25) aims to find a linear transformation that can maximize manifold margin and minimize the locality of all manifolds, simultaneously.

By using Lagrange multiplier, we can easily gain the solutions from the following eigen-decomposition equation:

$$S_MA_i = \lambda_iS_tA_i \tag{26}$$

Thus it can be concluded that A is composed of the eigenvectors related to d top eigenvalues of the generalized eigen-decomposition mentioned above. On one hand, the transformation matrix A is achieved by training samples. On the other hand, for any new coming test data X_t, its low dimensional embedding will be found by the linear transformation $Y_t = A^TX_t$. The outline of the proposed LLRMM is stated in Table 1.

3. Experiments and Discussions

In the following, we will carry out experiments to test the performance of the proposed LLRMM. Moreover, some related feature extraction methods including kernel principal component analysis (KPCA), nonparametric discriminant analysis (NDA), reconstructive discriminant analysis (RDA) and discriminant multiple manifold learning (DMML) are all involved in to make comparisons. In these comparison approaches, both RDA and LLRMM are of a Fisher form in their objective functions, which may easily lead to small sample size problem when applying them to real-world data. So PCA has to be adopted in advance, by which the dimensionality of the original data is reduced to avoid the problem. And then, KPCA, NDA, RDA, DMML and LLRMM are employed to handle the preprocessing data. At last, the nearest neighbor (NN) classifier is recommended to identify the extracted features by KPCA, NDA, RDA, DMML and LLRMM, respectively. In addition, as a baseline for nearest neighbor classifier, large margin nearest neighbor (LMNN) (Domeniconi

Table 1
Outline of the proposed LLRMM.

Input: The original data matrix $X = [X_1, X_2, \dots, X_n] \in \mathbb{R}^{D\times n}$ and their labels $C = [C_1, C_2, \dots, C_c]$, values of k, k is numbers of nearest neighbors when constructing the between-manifold graph, the within-manifold graph and the total-manifold graph.
Output: Projection matrix $A \in \mathbb{R}^{D\times d}$ and the low dimensional embeddings $Y = A^T X \in \mathbb{R}^{d\times n}$
Algorithm:
Step 1 Construction of the between-manifold graph, the within-manifold graph and the total-manifold graph
1.1 Determine k and construct the between-manifold graph, the within-manifold graph and the total-manifold graph with the fixed k;
1.2 Compute the optimal weight matrix W_w as Eq. (5), W_b as Eq. (6) and W_t as Eq. (7) , respectively.
Step 2 Eigen decomposition
2.1 Calculate matrices $U_w = (I - W_w)(I - W_w)^T, U_b = (I - W_b)(I - W_b)^T$ and $U_t = (I - W_t)(I - W_t)^T$, respectively;
2.2 Obtain the corresponding within-manifold graph scatter $S_w = XU_wX^T$, the between-manifold graph scatter $S_b = XU_bX^T$ and the total-manifold graph scatter $S_t = XU_tX^T$, respectively;
2.3 Compute manifold margin $S_M = S_b - S_w = X(U_b - U_w)X^T$;
2.4 Solve the generalized eigen-decomposition $S_M A = \lambda S_t A$;
2.5 Sort their eigenvectors $[A_1, A_2, \dots, A_d]$ according to their associated eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$.
Step 3 Low dimensional embeddings
3.1 Output the transformation matrix $A = [A_1, A_2, \dots, A_d]$
3.2 output the low dimensional embeddings for any new coming point X_t with the transformation matrix A, i.e. $Y_t = A^T X_t$

Fig. 2. One person images in AR face subset.

et al., 2005) is also introduced to make comparisons. However, because LMNN is a classifier, so we just compare its performance to those of the above mentioned feature extraction methods with NN classifier. At last all these methods are applied to some benchmark face data sets such as AR, CMU PIE, Yale, YaleB and LFW. The recognition performances will be individually reported according to the experiments on each face data set.

3.1. Experiments on AR face data

There are over 4,000 color face images of 70 men and 56 women in AR face data set, which contains frontal views of faces by changing facial expressions, lighting conditions and occlusions. Among them, 120 persons including 65 men and 55 women face images are selected, which can be divided into two sessions by taking photos on separated two weeks. Each session has 13 color images, which can also be transformed to gray images. In this experiment, 14 grayscale face images of 120 individuals are chosen, where each session contains 7 images. Each image is manually cropped and normalized to be size of 64 × 64 pixels. The images of one person in AR face data subset are shown in Fig. 2.

In this experiment, we will test the impacts of the number of training samples, where 5,6,7 and 8 images per-class in AR subset are selected as training and the corresponding 9,8,7 and 6 images as test. When exploring the local patch using k nearest neighbors criterion, k is often set to 4, 5, 6 and 7, respectively.

In the following, we illustrate the statistics results by using KPCA + NN, NDA + NN, RDA + NN, DMML + NN, LMNN and LLRMM + NN,

Table 2
Mean recognition rates with standard deviations on AR face data.

	5 trains	6 trains	7 trains	8 trains
DMML + NN	85.01 ± 1.24	87.35 ± 1.48	87.56 ± 1.62	90.43 ± 1.35
KPCA + NN	90.36 ± 1.53	91.76 ± 1.97	92.54 ± 1.28	93.52 ± 1.01
NDA + NN	92.45 ± 1.06	93.21 ± 1.12	94.03 ± 0.89	95.33 ± 1.47
RDA + NN	92.98 ± 1.74	93.99 ± 2.10	94.89 ± 1.02	96.21 ± 0.72
LMNN	84.62 ± 1.49	87.82 ± 1.43	90.29 ± 0.74	92.17 ± 1.29
LLRMM + NN	95.54 ± 1.32	96.44 ± 1.22	97.21 ± 0.99	98.01 ± 0.78

Table 3
Mean recognition rates with standard deviations on CMU PIE face data.

	60 trains	70 trains	80 trains	90 trains
DMML + NN	88.23 ± 2.12	89.79 ± 1.74	90.72 ± 1.22	91.03 ± 0.99
KPCA + NN	90.02 ± 1.66	91.57 ± 1.25	92.63 ± 1.18	93.26 ± 0.87
NDA + NN	91.34 ± 1.35	92.76 ± 1.62	93.29 ± 1.37	94.08 ± 1.35
RDA + NN	91.92 ± 1.54	93.42 ± 1.53	94.37 ± 0.98	95.01 ± 1.06
LMNN	92.29 ± 0.26	93.70 ± 0.34	94.99 ± 0.20	95.24 ± 0.17
LLRMM + NN	93.24 ± 1.68	94.38 ± 1.02	95.48 ± 0.77	95.99 ± 1.52

where experiments are repeated 10 times by randomly selecting 5,6,7 and 8 training samples each person on AR face subset, respectively. The mean recognition rates and their standard deviations are displayed in Table 2, from which we can find that the mean recognition rates of LLRMM + NN are superior to those of the others if the same training samples are selected.

3.2. Experiments on CMU PIE face data

The face image data set of CMU PIE is widely used, especially for pose, illumination and expression evaluation. In CMU PIE face data set, 41,368 face images belonging to 68 persons are contained. By varying pose, illumination, and expression, 13 synchronized cameras and 21 flashes were used to capture the face images. In our experiment, 170 gray-scale face images for each individual are chosen, which consist of the CMU PIE face subset. Moreover, each image is resized to be 32 × 32. An example images are shown in Fig. 3.

In the experiments, 60, 70, 80 and 90 images of one class are selected from the CPU PIE subset as training and the corresponding 110, 100, 90 and 80 samples for one person are taken as test. When constructing the within-manifold graph, the between-manifold graph and the total-manifold graph, k is set to 20.

Shown in Table 3 are statistics results including mean accuracies and the corresponding standard deviations in CMU PIE face subset, where we have conducted each experiment 10 times by randomly selecting personal 60, 70, 80 and 90 images as training. From Table 3, a conclusion can also be drawn that the mean recognition rates of LLRMM are larger than those of KPCA, NDA, RDA, DMML and LMNN with the same number of the training samples.

3.3. Experiments on Yale face data

Yale face database has been collected and constructed by the Yale Center for Computation Vision and Control. In the Yale face data set, 15 individuals have 165 images with variations in lighting condition, facial expression and with or without glasses, etc. Thus for each person, he(or she) has 11 images. Fig. 4 shows one person images from Yale face data set, which have been cropped to be of size 64 × 64 pixels.

The experiments on Yale face data set are also repeated 10 times to obtain the mean accuracies and the corresponding standard deviations. For each feature extraction method, the number of personal training data is set to 4, 5 and 6 and the rest as test samples. Meanwhile, k is also set to 3,4 and 5 when using k nearest neighbor criterion to construct the corresponding graphs, respectively. Table 4 shows the mean recognition rates and the corresponding standard deviations for KPCA + NN, NDA + NN, RDA + NN, DMML + NN, LMNN and LLRMM + NN with different training samples per class, where the proposed algorithm gains the best results.

Fig. 3. One person images from CMU PIE face subset.

Fig. 4. One person images Yale face data set.

Table 4
Mean recognition rates with standard deviations on Yale face data.

	4 trains	5 trains	6 trains
DMML + NN	81.35 ± 1.64	86.38 ± 1.43	88.94 ± 1.86
KPCA + NN	82.73 ± 1.57	90.12 ± 1.25	92.76 ± 0.89
NDA + NN	84.78 ± 1.02	91.56 ± 1.78	93.87 ± 2.21
RDA + NN	87.76 ± 0.92	92.18 ± 1.56	94.59 ± 1.36
LMNN	87.86 ± 2.14	89.86 ± 3.47	91.36 ± 2.20
LLRMM + NN	89.14 ± 0.85	93.33 ± 1.36	95.73 ± 1.11

Table 5
Mean recognition rates with standard deviations on YaleB face data.

Methods	20 trains	30 trains	40 trains
DMML + NN	72.13 ± 1.78	86.75 ± 2.31	89.57 ± 2.13
KPCA + NN	73.24 ± 2.56	88.38 ± 2.78	90.35 ± 1.72
NDA + NN	76.89 ± 1.98	89.78 ± 1.96	91.48 ± 1.64
RDA + NN	78.87 ± 1.35	90.88 ± 1.84	92.53 ± 1.47
LMNN	79.73 ± 0.76	91.38 ± 1.11	93.51 ± 2.01
LLRMM + NN	82.12 ± 2.34	92.57 ± 1.04	95.27 ± 1.45

3.4. Experiment on YaleB face data

The YaleB face data set is an extension to the original Yale face database, where 28 persons with 16 128 images are collected under 9 poses and 64 illumination conditions, where we can select 38 individuals. Furthermore, 64 near frontal images under varied illuminations for each person are selected as a YaleB face data subset in our experiments. Every image in the subset is cropped to be 32 × 32 pixels. Some images of one person in the subset are displayed in Fig. 5.

When carrying out the experiments, we select 20, 30 and 40 images per person as training samples and the rest 44, 34 and 24 samples as test, respectively. Moreover, k is set to 12 when determining k nearest neighbors. By repeating each experiment 10 times, the statistical results will be obtained, which are shown in Table 5. Moreover, it can also find that no matter how many training samples per class are selected, the statistics performances of the proposed LLRMM with NN classifier are also on the top compared to other methods such as DMML + NN, KPCA + NN, NDA + NN, RDA + NN, and LMNN classifier.

3.5. Experiment on LFW face data

In order to study unconstrained images for identity verification and face recognition, the Labeled Faces in the Wild (LFW) face database is constructed, where more than 13,000 face images from 1680 persons pictured under the unconstrained conditions are contained. In this experiment, we use a subset including 1251 images from 86 peoples and each has only 10–20 images (Wang et al., 2012). Each face image was manually cropped to be size of 32 × 32 pixels. Some face images for one person in the LFW subset are shown in Fig. 6.

In the experiments, we randomly select 6, 7 and 8 images of each subject as training samples and the rest as test samples. Moreover, k is respectively set to 5,6 and 7 when 6, 7 and 8 images per class are chosen as training samples. By repeating the experiment ten times, the statistical results are shown in Table 6. Because the LFW database is a very difficult database for image classification, the accuracies obtained by using different comparison methods are comparatively not high, which can also be found from Table 6. However, the mean recognition rates of the proposed method are superior to those of the other comparison ones.

Table 6
Mean recognition rates with standard deviations on LFW face data.

Methods	6 trains	7 trains	8 trains
DMML + NN	29.88 ± 1.73	32.26 ± 1.45	33.81 ± 1.95
KPCA + NN	31.52 ± 0.94	34.45 ± 1.76	35.27 ± 2.18
NDA + NN	32.83 ± 1.54	35.03 ± 1.99	36.28 ± 1.96
RDA + NN	33.36 ± 1.94	36.30 ± 1.58	37.21 ± 1.43
LMNN	34.56 ± 1.25	36.24 ± 2.12	37.56 ± 2.02
LLRMM + NN	36.58 ± 1.34	38.12 ± 1.72	40.27 ± 1.57

3.6. Discussions

From the above experimental results, we find that data scale shows impacts on the data recognition performance. In the experiments, some large data sets including AR, CMU PIE, YaleB and LFW as well as small data set such as Yale have been taken to test the proposed method. Experimental results on AR, CMU PIE, YaleB and LFW show that LLRMM + NN outperforms DMML + NN, KPCA + NN, NDA + NN, RDA + NN and LMNN. For large data sets, more data points are involved in construction of the between-manifold graph and the within-manifold graph. Thus the locality of manifolds can be well explored and the supervised information can be fully taken advantage of, which make contributions to data classification. For Yale data, although mean recognition rates using LLRMM + NN are still larger than those using DMML + NN, KPCA + NN, NDA + NN, RDA + NN and LMNN when the same number training samples are selected. However, the mean accuracies of the proposed LLRMM + NN show no much superiority to those of DMML + NN, KPCA + NN, NDA + NN, RDA + NN. The reason lies in that LLRMM, DMML, NDA and RDA are manifold learning based methods, which heavily depend on construction of k nearest neighbor graph, if the training samples are too small to mine the manifold structure, the corresponding methods cannot gain better results.

4. Conclusions and future work

In this paper, LLRMM is presented for dimensionality reduction or feature extraction, where a novel manifold margin is globally defined. In the proposed LLRMM, manifold label information is taken advantage of to guide the construction of a between-manifold graph and a within-manifold graph respectively. In the within-manifold graph, any node and its neighborhood points must be resided on one manifold. On the contrary, in the between-manifold graph, any node and its k nearest

Fig. 5. Images for one person in YaleB face subset.

Fig. 6. Images for one person in LFW face subset.

neighbors should be located on different manifolds. On the basis of the between-manifold graph and the within-manifold graph, a manifold margin is newly modeled, which can be reasoned to difference of the minimum representation errors in the between-manifold graph to those in the within-manifold graph. At last, it will explore a subspace by maximizing the proposed manifold margin and minimizing the manifold locality, simultaneously. By making comparisons to some related pattern recognition approaches as KPCA + NN, NDA + NN, RDA + NN, DMML + NN and LMNN, experiments results on AR, CMU PIE, Yale, YaleB and LFW face data show that LLRMM + NN outperforms them. But when constructing both the between-manifold graph and the within-manifold graph in the proposed LLRMM, the parameter, i.e. k, which has impacts on determining the linearly local patch on the manifold, are set to be the same. In the future, it will be expected to conduct some works to explore the classification performance by setting different k in both graphs.

Acknowledgments

This work was partly supported by grants of National Natural Science Foundation of China (61572381, 61273303, 61472280 and 61403287) and Post-doctoral Science Foundation of China (2016M601646).

References

Belkin, M., Niyogi, P., 2003. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. 15 (6), 1373–1396.

Cai, D., He, X., Zhou, K., Han, J., Bao, H., 2007. Locality Sensitive Discriminant Analysis. In: 20th international joint conference on Artificial intelligence, pp. 708-713.

Chen, H.T., Chang, H.W., Liu, T.L., 2005. Local discriminant embedding and its variants. In: 2005 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 846–853.

Chen, Y., Jin, Z., 2012. Reconstructive discriminant analysis: A feature extraction method induced from linear regression classification. Neurocomputing 87, 41–50.

Domeniconi, C., Gunopulos, D., Peng, J., 2005. Large margin nearest neighbor classifiers. IEEE Trans. Neural Netw. 16 (4), 899–909.

Donoho, D., Grams, C., 2003. Hessian eigenmaps: Locally linear embedding techniques for high dimensional data. Proc. Natl Acad. Sci. 100, 5591–5595.

Han, P.Y., Beng, A.T.J., Kiong, W.E., 2008. Neighborhood discriminant locally linear embedding in face recognition. In: International Conference on Computer Graphics, Imaging and Visualization, pp. 223-228.

Huang, D.S., 2004a. A constructive approach for finding arbitrary roots of polynomials by neural networks. IEEE Trans. Neural Netw. 15 (2), 477–491.

Huang, D.S., 2004b. Systematic Theory of Neural Networks for Pattern Recognition (in Chinese). Publishing House of Electronic Industry of China.

Huang, D.S., Chi, Z.R., Siu, W.C., 2005. A case study for constrained learning neural root finders. Appl. Math. Comput. 165 (3), 699–718.

Huang, D.S., Du, J.X., 2008. A constructive hybrid structure optimization methodology for radial basis probabilistic neural networks. IEEE Trans. Neural Netw. 19 (12), 2099–2115.

Hui, K., Wang, C., 2008. Clustering-based locally linear embedding. In: International Conference on Pattern Recognition, pp. 2054-2056.

Jolliffe, I.T., 2002. Principal Component Analysis. Springer.

Kim, T.K., Stenger, B., Kittler, J., Cipolla, R., 2011. Incremental linear discriminant analysis using sufficient spanning sets and its applications. Int. J. Comput. Vis. 91 (2), 216–232.

Lawrence, K.S., 2001. An Introduction to Locally Linear Embedding. URL: <http://www.cs.toronto.edu/~roweis/lle/>.

Li, Z., Lin, D., Tang, X., 2009. Nonparametric discriminant analysis for face recognition. IEEE Trans. Pattern Anal. Mach. Intell. 31 (4), 755–761.

Li, B., Wang, Y.Q., Lei, L., Fan, Z.T., 2015. Locally linear representation manifolds margin. In: International Conference on Intelligent Computing, pp. 483-490.

Li, B., Zheng, C., Huang, D., 2008. Locally linear discriminant embedding: an efficient method for face recognition. Pattern Recognit. 41 (12), 3813–3821.

Lin, T., Zha, H.B., 2008. Riemannian manifold learning. IEEE Trans. Pattern Anal. Mach. Intell. 30 (5), 796–809.

Lu, J., Tan, Y.P., Wang, G., 2013. Discriminative multimanifold analysis for face recognition from a single training sample per person. IEEE Trans. Pattern Anal. Mach. Intell. 35 (1), 39–51.

van der Maaten, L.J.P., 2014. Accelerating t-SNE using tree-based algorithms. J. Mach. Learn. Res. 15, 3221–3245.

Martinez, A.M., Kak, A.C., 2001. PCA versus LDA. IEEE Trans. Pattern Anal. Mach. Intell. 23 (2), 228–233.

Motta, R., Minghim, R., Lopes, A.D.A., Oliveira, M.C.F., 2015. Graph-based measures to assist user assessment of multidimensional projections]. Neurocomputing 150, 583–598.

Pang, Y., Liu, Z., Yu, N., 2006. A new nonlinear feature extraction method for face recognition. Neurocomputing 69, 949–953.

de Ridder, D., Kouropteva, O., Okun, O., 2003. Supervised locally linear embedding. In: International Conference on Artificial Neural Networks, pp. 333–341.

de Ridder, D., Loog, M., Reinders, M.J.T., 2004. Local Fisher embedding. In: International Conference on Pattern Recognition, pp. 295–298.

Roweis, S.T., Saul, L.K., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290, 2323–2326.

Sadatnejad, K., Ghidary, S.S., 2016. Kernel learning over the manifold of symmetric positive definite matrices for dimensionality reduction in a BCI application]. Neurocomputing 179, 152–160.

Saul, L.K., Roweis, S.T., 2003. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. J. Mach. Learn. Res. 4, 119–155.

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

Shang, L., Huang, D.S., Du, J.X., Zheng, C.H., 2006. Palmprint recognition using Fast ICA algorithm and radial basis probabilistic neural network. Neurocomputing 69 (13–15), 1782–1786.

Sugiyama, M., 2006. Local Fisher Discriminant Analysis for Supervised Dimensionality Reduction. In: International Conference on Machine Learning, pp. 905-912.

Sun, Z., Lam, K.M., Gao, Q.W., 2013. Depth estimation of face images using the nonlinear least squares model. IEEE Trans. Image Process. 22 (1), 17–30.

Tenenbaum, J.B., de Silva, V., Langford, J.C., 2000. A global geometric framework for nonlinear dimensionality reduction. Science 290, 2319–2323.

Wang, D., Shen, H., Truong, Y., 2016. Efficient dimension reduction for high-dimensional matrix-valued data. Neurocomputing 190, 25–34.

Wang, S.J., Yang, J., Sun, M.F., Peng, X.J., Sun, M.M., Zhou, C.G., 2012. Sparse tensor discriminant color space for face Verification. IEEE Trans. Neural Netw. Learn. Syst. 23 (6), 876–888.

Weinberger, K.Q., Saul, L.K., 2006. An introduction to nonlinear dimensionality reduction by maximum variance unfolding. In: National Conference on Artificial Intelligence, pp. 1683-1686.

Wen, G., Jiang, L., 2006. Clustering-based locally linear embedding. In: 2006 IEEE International Conference on Systems, Man and Cybernetics. IEEE, pp. 4192–4196.

Wen, Y., He, L., Shi, P., 2012. Face recognition using difference vector plus KPCA. Digit. Signal Process. 22, 140–146.

Xu, D., Yan, S., Tao, D., Lin, S., Zhang, H.J., 2007. Marginal fisher analysis and its variants for human gait recognition and content-based image retrieval. IEEE Trans. Image Process. 16 (11), 2811–2820.

Yan, S., Xu, D., Zhang, B., Zhang, H.J., 2007. Graph embedding: A general framework for dimensionality reduction. IEEE Trans. Pattern Anal. Mach. Intell. 29 (1), 40–51.

Yang, J., Jin, Z., Yang, J., Zhang, D., Frangi, A.F., 2004. Essence of kernel fisher discriminant: KPCA plus LDA. Pattern Recognit. 37 (10), 2097–2100.

Yang, J., Lou, Z., Jin, Z., Yang, J., 2008. Minimal local reconstruction error measure based discriminant feature extraction and classification. In: Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1–6.

Yang, H., Zhang, C., 2008. The research on two kinds of restricted biased estimators based on mean squared error matrix. Commun. Stat. : Theory Methods 37 (1), 70–80.

Yu, Q., Wang, R., Yang, X., Li, B.N, Yao, M., 2016. Diagonal principal component analysis with non-greedy ℓ1-norm maximization for face recognition. Neurocomputing 171, 57–62.

Zhang, H., Gao, X., Wu, P., Xu, X., 2016a. A cross-media distance metric learning framework based on multi-view correlation mining and matching. World Wide Web 19 (2), 181–197.

Zhang, J., Li, H., Zhou, Z.H., 2006. Ensemble-based discriminant manifold learning for face recognition. In: International Conference on Computing, Networking and Communication. pp. 29–38.

Zhang, J., Shen, H., Zhou, Z.H., 2004. Unified locally linear embedding and linear discriminant analysis algorithm for face recognition. Lecture Notes in Comput. Sci. 3338, 296–304.

Zhang, H., Wu, P., Beck, A., Zhang, Z.J., Gao, X.Y., 2016b. Adaptive incremental learning of image semantics with application to social robot. Neurocomputing 173, 93–101.

Zhang, Z., Zha, H., 2005. Principal manifolds and nonlinear dimension reduction via local tangent space alignment. SIAM J. Sci. Comput. 26 (1), 313–338.

Zhang, H., Zhang, W., Liu, W., Xu, X., Fan, H., 2016c. Multiple kernel visual-auditory representation learning for retrieval. Multimedia Tools Appl. 75, 9169–9184.

Zhang, Z., Zhao, L., 2007. Probability-based locally linear embedding for classification. In: International Conference on Fuzzy Systems and Knowledge Discovery, pp. 243–247.

Zhao, L., Zhang, Z., 2009. Supervised locally linear embedding with probability-based distance for classification. Comput. Math. Appl. 57 (6), 919–926.

Zhao, Z.Q., Huang, D.S., Sun, B.Y., 2004. Human face recognition based on multiple features using neural networks committee. Pattern Recognit. Lett. 25 (12), 1351–1358.

Zhao, Q., Zhang, D., Lu, H., 2005. Supervised LLE in ICA space for facial expression recognition. In: International Conference on Neural Networks Brain, pp. 1970–1975.