Research Article

Maximum Neighborhood Margin Discriminant Projection for Classification

Jianping Gou, Yongzhao Zhan, Min Wan, Xiangjun Shen, Jinfu Chen, and Lan Du

1 School of Computer Science and Telecommunication Engineering, JiangSu University, ZhenJiang, JiangSu 212013, China
2 School of Mathematics and Computer Engineering, Xihua University, Chengdu, Sichuan 610039, China
3 Department of Computing, Macquarie University, Sydney, NSW 2109, Australia

Correspondence should be addressed to Jianping Gou; cherish.gjp@gmail.com and Yongzhao Zhan; yzzhan@ujs.edu.cn

Received 16 December 2013; Accepted 8 January 2014; Published 20 February 2014

Academic Editors: P. Bala and C.-W. Tsai

Copyright © 2014 Jianping Gou et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

We develop a novel maximum neighborhood margin discriminant projection (MNMDP) technique for dimensionality reduction of high-dimensional data. It utilizes both the local information and class information to model the intraclass and interclass neighborhood scatters. By maximizing the margin between intraclass and interclass neighborhoods of all points, MNMDP cannot only detect the true intrinsic manifold structure of the data but also strengthen the pattern discrimination among different classes. To verify the classification performance of the proposed MNMDP, it is applied to the PolyU HRF and FKP databases, the AR face database, and the UCI Musk database, in comparison with the competing methods such as PCA and LDA. The experimental results demonstrate the effectiveness of our MNMDP in pattern classification.

1. Introduction

Dimensionality reduction (DR) plays an important role in many fields such as pattern classification, machine learning, and computer vision. Its purpose is to solve the “curse of dimensionality” [1] and map the high-dimensional points to a subspace that reveals the intrinsic structure of the original data. Among the DR techniques, principal component analysis (PCA) [2, 3], linear discriminant analysis (LDA) [3, 4], locality preserving projection (LPP) [5, 6], and their kernelized and tensorized variants [7–18] are the most representative and well-known algorithms. Although these methods have different suppositions, they can be put into a unified graph embedding framework with different constraints [19].

Generally, PCA aims to preserve the global geometric structure of data by maximizing the trace of the feature covariance matrix and produces compact representation of the original space in a low-dimensional space. However, it does not take into account the class label information. The goal of LDA is to find the global discriminant information for classification by maximizing the ratio between interclass and intraclass scatters. In contrast to PCA, LDA takes much consideration of the class information and strengthens the ability of pattern discrimination. Since both PCA and LDA only consider the global structure of data, they have little to do with the essential manifold of the data. As for them, it is difficult to discover the hidden submanifold that truly reflects the essential structure of the data.

In contrast to PCA and LDA, locality preserving projection (LPP) is introduced to detect the intrinsic geometry of the manifold structure of data [5, 6]. LPP is a classical linear graph embedding [19] derived from Laplacian Eigenmap [20]. It attempts to find an embedding that preserves the local neighborhood information and reflects the inherent submanifold structure. In recent years, many variants of LPP have been developed for dimensionality reduction [21–34]. These LPP-based DR algorithms can be mainly put into two categories: supervised and unsupervised ones. As for the supervised LPP-based methods, they generally employ class information of data in the process of subspace learning, such as in [23–27, 29–32]. On the contrary, the unsupervised LPP-based ones do not consider class information [21, 22, 33, 34]. Among them, unsupervised discriminant projection (UDP) [21], as a simplified version of LPP [35], is a very popular
method with the aim of resolving the “overlearning locality”
existing in LPP. UDP seeks to find a projection by maximizing
the nonlocal scatter and minimizing the local scatter simulta-
neously. As for LPP and UDP, since the local structure of
data is modelled by the nearest neighbor graph, they may
be not effective in the case of noisy data. Moreover, they are
completely unsupervised in regard to the class label informa-
tion and cannot perform well for classification. In addition,
recently there are some new DR techniques that integrate
sparse representation (SR) and subspace learning [36–41].
Nowadays, locality preserving projection (LPP) [5, 6] is
a classical graph embedding [19]. Through a liner transforma-
tion projection, LPP can find an embedding that best
discovers the intrinsical manifold structure of data. In order
to do so, it first constructs a weighted affinity graph
\( G = (\mathcal{V}, \mathcal{E}) \), where \( \mathcal{V} \) is the set of all points of data and \( \mathcal{E} \) is
the set of edges between any pairs of points. Note that the affinity
graph \( G \) is usually established by \( k \)-neighborhood; nodes \( i \) and
\( j \) are connected by an edge if \( x_i \) is in \( k \) nearest neighbors of
\( x_j \) or \( x_j \) is in \( k \) nearest neighbors of \( x_i \). Then, the weight
\( W_{ij} \) of an edge between \( x_i \) and \( x_j \) is often defined as follows:
\[
W_{ij} = \begin{cases} 
\exp \left( -\frac{\| x_i - x_j \|^2}{\delta} \right), & x_j \in \mathcal{N}_k(x_i) \text{ or } x_i \in \mathcal{N}_k(x_j) \\
0, & \text{otherwise},
\end{cases}
\]
where \( \mathcal{N}_k(x_i) \) or \( \mathcal{N}_k(x_j) \) denotes a set of the \( k \) nearest
neighbors of the sample \( x_i \) or \( x_j \) and the parameter \( \delta \) is
a positive constant.

To obtain a liner graph embedding, the objective function of
LPP to be minimized is as follows:
\[
\min \sum_{ij} \| y_i - y_j \|^2 W_{ij}.
\]
The criterion above with \( W_{ij} \) can give a penalty for mapping
eighboring points \( x_i \) and \( x_j \) far apart in an embedded sub-
space. By simple algebra, the objective criterion for LPP under
the appropriate constraint can be rewritten as:
\[
\begin{align*}
\underset{\Phi}{\text{arg min}} \quad & \text{tr}\left( \Phi^T X L X^T \Phi \right) \\
\text{s.t.} \quad & \Phi^T X D X^T \Phi = I,
\end{align*}
\]
where \( D \) is a diagonal matrix and its entries are column (or
row, since \( W \) is symmetric) sum of \( W \); that is, \( D_{ij} = \sum_j W_{ij} \)
and \( L = D - W \) are the Laplacian matrix. In (4), \( \text{tr}(\cdot) \) and \( I \)
denote the trace of the matrix and the identity matrix,
respectively.

Finally, the transformation matrix \( \Phi \) to minimize (4)
can be achieved by solving the generalized eigenvalue problem
\[
X L X^T \Phi = \lambda X D X^T \Phi,
\]
where \( \Phi \) only contains \( d \) eigenvectors corresponding to the \( d \)
smallest eigenvalues, that is, \( \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d \).
3. Maximum Neighborhood Margin Discriminant Projection

In pattern recognition, the class label information of data plays an important role for classification. Moreover, the local structures of the training samples are also very useful for it. Inspired by the two facts, a novel dimensionality reduction method, called maximum neighborhood margin discriminant projection (MNMDP), is introduced in this section. In the proposed MNMDP, both the class information and local geometry of data are fully taken into account for classification simultaneously. The major focus of MNMDP is to seek a nonlinear projection (MNMDP), is introduced in this section. In this method, called maximum neighborhood margin discrimination, the local neighborhood structure and the intraclass discriminating weight, that is, 1 + \( \exp\left(-\frac{\|x_i - x_j\|^2}{\delta}\right) \), can represent the class information of the same classes. From (6), the affinity weights of the intraclass points are larger than those in LPP. This fact is very advantageous to classification.

From the viewpoint of pattern recognition, it is quite favorable for presuming that the different samples have different contributions to classification. Generally speaking, the fact that the samples with greater contributions have the more significance for classification is naturally related to their neighborhood location in the feature space. Here, we take into account a local scaling regulator of data to dynamically adjust adjacent weights between pairs of neighbors, so as to reasonably reflect the classification contribution of each sample. According to the k-neighbors of one sample \( x_i \), the parameter \( \delta \) as local scaling regulator in (6) is set to be as follows:

\[
\delta = \frac{1}{k^2} \sum_{j=1}^{k} \| x_i - x_j \|^2.
\]

This can be a good reasonable way to estimate the value of \( \delta \), and the affinity weights \( W_{ij}^+ \) between nodes \( i \) and \( j \) are allowed to self-tune in terms of the k-neighbors with \( \delta \).

To still retain the intraclass neighborhood relations through a linear mapping, that is, \( y_j = \Phi^T x_i \), the intraclass neighborhood scatter along a projection \( \phi \) is defined as

\[
J^*(\phi) = \frac{1}{2} \sum_{ij} \| y_i - y_j \|^2 W_{ij}^+.
\]

It follows from (8) that

\[
J^*(\phi) = \frac{1}{2} \sum_{ij} \| \phi^T x_i - \phi^T x_j \|^2 W_{ij}^+
\]

\[= \phi^T \left( \frac{1}{2} \sum_{ij} (x_i - x_j) (x_i - x_j)^T W_{ij}^+ \right) \phi \]

\[= \phi^T S^+ \phi,
\]

where \( S^+ \) is called the intraclass neighborhood scatter matrix:

\[
S^+ = \frac{1}{2} \sum_{ij} (x_i - x_j) (x_i - x_j)^T W_{ij}^+.
\]

To gain more insight into (10) in terms of the affinity matrix \( W^+ \) from (6), \( S^+ \) is rewritten as

\[
S^+ = \frac{1}{2} \left( \sum_{ij} W_{ij}^+ x_i x_i^T + \sum_{ij} W_{ij}^+ x_i x_j^T - 2 \sum_{ij} W_{ij}^+ x_i x_j \right)
\]

\[= \sum_i D_i^+ x_i x_i^T - \sum_{ij} W_{ij}^+ x_i x_j^T \]

\[= XD^+ X^T - XW^+ X^T \]

\[= XL^+ X^T, \]

where \( D^+ \) is a diagonal matrix, its elements are column sum of \( W^+ \), that is, \( D_{ii}^+ = \sum_{ij} W_{ij}^+ \) and \( L^+ = D^+ - W^+ \). Note that \( L^+ \), \( D^+ \), and \( S^+ \) are symmetric matrices.
In order to well preserve the intraclass neighborhood and keep the intraclass neighborhood scatter compact in the embedded subspace, the optimal projections can be obtained by minimizing the intraclass neighborhood scatter:

\[
\text{Min } J^+ (\phi) = \phi^T S^+ \phi. \tag{12}
\]

3.2. Interclass Neighborhood Scatter. In contrast to intraclass neighborhood scatter, the affinity weights for interclass neighborhoods of all points are defined as follows:

\[
W_{ij} = \begin{cases} 
\exp \left( -\frac{\|x_i - x_j\|^2}{\delta} \right) \left( 1 - \exp \left( -\frac{\|x_i - x_j\|^2}{\delta} \right) \right), & x_i \in \mathcal{N}_k^+(x_j) \text{ or } x_j \in \mathcal{N}_k^-(x_i), \\
0, & \text{otherwise},
\end{cases} \tag{13}
\]

where the parameter \(\delta\) is a positive regulator, the same as (7). In (13), the affinity weight can simultaneously reflect the local interclass neighborhood structure by the local weight \(\exp(-\|x_i - x_j\|^2/\delta)\) and the class information of the different classes by interclass discriminating weight \(1 - \exp(-\|x_i - x_j\|^2/\delta)\). From (13), the fact that the affinity weights of the interclass points are less than those in LPP is also very helpful for classification.

Then, the interclass neighborhood scatter along a projection \(\phi\) is defined as

\[
J^- (\phi) = \frac{1}{2} \sum_{ij} \| y_i - y_j \|^2 W_{ij}^-.
\]

It follows from (14) that

\[
J^- (\phi) = \frac{1}{2} \sum_{ij} \| \phi^T x_i - \phi^T x_j \|^2 W_{ij}^- = \phi^T S^- \phi,
\]

where \(S^-\) is called the interclass neighborhood scatter matrix:

\[
S^- = \frac{1}{2} \sum_{ij} (x_i - x_j)(x_i - x_j)^T W_{ij}^-.
\]

By the same algebra as \(S^+, S^-\) in (16) is rewritten as follows:

\[
S^- = X L^- X^T,
\]

where \(D^-\) is a diagonal matrix, its elements are column sum of \(W^-\), that is, \(D^-_{ii} = \sum_j W_{ij}^-\), and \(L^- = D^- - W^-\). Note that \(L^+, D^+,\) and \(S^-\) are symmetric matrices.

To gain more discrimination between different classes through a linear mapping, the interclass neighborhood scatter in the projected subspace should be kept more separable by maximizing the following criterion:

\[
\text{Max } J^- (\phi) = \phi^T S^- \phi. \tag{18}
\]

3.3. Optimal Linear Embedding. Combining (12) and (18) with the orthonormal constraint (i.e., \(\Phi^T \Phi = I\)), we get the following objective function:

\[
\begin{align*}
\text{Min } & J^+ (\Phi) = \text{tr} (\Phi^T S^+ \Phi) \\
\text{Max } & J^- (\Phi) = \text{tr} (\Phi^T S^- \Phi) \tag{19}
\end{align*}
\]

s.t. \(\Phi^T \Phi = I\),

where \(\Phi = [\phi_1, \phi_2, \ldots, \phi_d]\) and \(\phi_i\) is an orthogonal vector. Based on the idea of the maximum margin criterion [43], (19) can be reformulated as follows:

\[
\begin{align*}
\arg \max_{\Phi} & \quad \text{tr} (\Phi^T (S^- - S^+) \Phi) \tag{20} \\
\text{s.t.} & \quad \Phi^T \Phi = I,
\end{align*}
\]

or

\[
\begin{align*}
\max & \quad \sum_{i=1}^d \phi_i^T (S^- - S^+) \phi_i \\
\text{s.t.} & \quad \phi_i^T \phi_i = 1, \quad \phi_i^T \phi_j = 0 \quad (i \neq j).
\end{align*}
\]

According to (20) or (21), we can find two aspects that are favorable for classification. On one hand, the optimal projections obtained are such that the intraclass samples are attracted being more compact (minimizing the intraclass neighborhood scatter), while the interclass samples are simultaneously pulled being more separable (maximizing the interclass neighborhood scatter). Of course, it can keep the margin between intraclass and interclass neighborhood maximum in a new subspace, so as to clearly enhance pattern discrimination. On the other hand, the graph embedding obtained with orthogonal projections can have both more locality preserving power and more discriminating power [33].

To maximize the above objective function, we can use the Lagrangian multiplier method to first build the following function:

\[
L(\phi_i, \lambda_i) = \sum_{i=1}^d \left( \phi_i^T (S^- - S^+) \phi_i - \lambda_i (\phi_i^T \phi_i - 1) \right), \tag{22}
\]

where \(\lambda_i (i = 1, \ldots, d)\) is a Lagrange multiplier. Then, the optimization is carried out by the partial derivative of \(L(\phi_i, \lambda_i)\) with respect to \(\phi_i\):

\[
\frac{\partial L(\phi_i, \lambda_i)}{\partial \phi_i} = (S^- - S^+) \phi_i - \lambda_i \phi_i. \tag{23}
\]

Let (23) be zero; we yield

\[
(S^- - S^+) \phi_i = \lambda_i \phi_i. \tag{24}
\]

Thus, the optimal matrix \(\Phi\) that maximizes the objective criterion in (20) can be achieved by solving the generalized eigenvalue problem

\[
(S^- - S^+) \Phi = \lambda \Phi, \tag{25}
\]
In this section, we evaluate the classification performance of the proposed MNMDP, in comparison with the state-of-the-art DR methods: PCA, LDA, LPP, and UDP. The experiments are conducted on four databases including HRF, FKP, AR, and Musk databases. In order to measure the recognition performance, l images per class on each database are randomly selected as training samples, while the remaining images per class are used for testing. To ensure that the performance is not biased from certain random realization of training/testing set, 10 independent runs for a given l are performed. Since training and testing samples are chosen randomly online, those 10 runs give us 10 different training and testing sets for performance evaluation. Note that both training and testing sets on each run have no overlap between them. The average recognition rates across these 10 runs with 95% confidence are regarded as the final classification results. In our experiments, to overcome the small sample size problem encountered in LDA, LPP, and UDP, PCA preprocessing is employed to solve the singular matrix by keeping nearly 98% image energy. If the singularities still exist, pseudoinverse is adopted. In order to ensure a fair comparisons, the regulator $\delta$ in LPP is set in terms of (7). The $k$-neighborhood parameter denoted by $W_k$ for building the weight matrix in LPP and UDP is selected as $W_k = l - 1$ [21]. The nearest neighbor (NN) classifier with Euclidean distance is used for classification in the experiments.

4.1. Experiments on HRF. The PolyU HRF (High-Resolution-Fingerprint) database [44] was collected in two separate sessions. Here, we use the DBII of HRF. The database contains 148 fingers, each of which has five sample images per session. Each image is taken with 1200 dpi and the size per image is $640 \times 480$ pixels, with 256 grey levels per pixel. For computational efficiency, each image is resized to $32 \times 32$ pixels in our experiments. As an example, Figure 1 shows ten images of one finger in the HRF database. We form the training set by a random subset of $l$ images per class and use the rest as a testing set. In the following experiments, the numbers of training sample images per class are chosen as $l = 5, 6, 7, 8$.

In the experiments on HRF, we first explore the performance of MNMDP with varying $k$-neighborhood parameter $W_k$ in terms of recognition rates. The value of $W_k$ is set from 1 to 21 in Step 2. The maximal average classification results via $W_k$ for each $l$ are plotted in Figure 2. It can be obviously observed that the proposed MNMDP with more training samples has better classification. As can be seen in Figure 2, the performance of MNMDP for each $l$ first increases slowly when $W_k$ changes from 1 to 5, and then increases rapidly when $W_k$ changes from 5 to 9, and finally drops with increase of $W_k$. The possible reason for this experimental phenomenon is that the affinity graph is unable to capture effectively the geometry of data when $W_k$ is small and

---

**Algorithm 1**: Maximum neighborhood margin discriminant projection.

Require: $X \in \mathbb{R}^m$: a testing point. $X = \{x_i \in \mathbb{R}^m\}_{i=1}^{N}$: a training set.

Ensure: Predict the class label of $X$.

Step 1. Construct the adjacent graph $\mathcal{G}$ for any point in the training set $X$ using $k$-neighborhood.

Step 2. Compute the affinity weight matrix $W^+$ for intraclass neighborhood and $W^-$ for interclass neighborhood of any point, respectively.

If $x_j \in N_k^+(x_i)$ or $x_j \in N_k^-(x_i)$, then

$$W_{ij} = \exp \left( -\frac{\|x_i - x_j\|^2}{\delta} \right) \left( 1 + \exp \left( -\frac{\|x_i - x_j\|^2}{\delta} \right) \right)$$

else

$$W_{ij} = 0$$

end if

if $x_j \in N_k^+(x_i)$ or $x_j \in N_k^-(x_i)$, then

$$W_{ij} = \exp \left( -\frac{\|x_i - x_j\|^2}{\delta} \right) \left( 1 - \exp \left( -\frac{\|x_i - x_j\|^2}{\delta} \right) \right)$$

else

$$W_{ij} = 0$$

end if

Step 3. Compute the intraclass neighborhood scatter matrix $S^I = X^I X^T$ and the interclass neighborhood scatter matrix $S^O = X^O X^T$.

Step 4. Obtain the optimal projection matrix $\Phi$ by maximizing the generalized eigenvalue problem $(S^I - S^O)\Phi = \lambda \Phi$.

Step 5. Dimensionality reduction: transform all the points from the high-dimensional feature space $X$ to a subspace $Y$ with the optimized projection matrix $\Phi$, that is, $Y = \Phi^T X$.

Step 6. Classify $\overline{X}$ using a certain classifier. The projection of $\overline{X}$ is first obtained by $\overline{Y} = \Phi^T \overline{X}$ and then classify $\overline{Y}$ in the projected subspace $Y$.

---

Figure I: Ten sample images of one subject in the HRF database.
the more geometrical information of data can be preserved as \( W_k \) increases. However, when \( W_k \) is beyond the reasonable value, the \( k \)-neighborhood for a given point may include more interclass points [45], and this can degrade the ability of pattern discrimination. Consequently, the experimental results reveal that \( k \)-neighborhood parameter \( W_k \) in MNMDP plays an important role for preservation of geometrical and discriminant information of data that is available for classification and its suitable value can be easily selected to achieve good performance.

Furthermore, the experimental comparisons of competing methods are studied by varying the reduced dimensionality on HRF. We experiment with the dimension of the reduced space from 5 to 100 in Step 5. Based on the results shown in Figure 2 the best \( k \)-neighborhood parameters for MNMDP are set as \( W_k = 9 \) for \( l = 5, 6, 7, 8 \), respectively. For each \( l \) on HRF, the comparative average recognition performance of each method is given in Figure 3. As shown in Figure 3, the classification performance of each method ascends quickly until the dimensionality is about 30 and then keeps almost stable or decreases slowly with increase of dimensionality. It is clear that the proposed MNMDP consistently outperforms the other methods at any value of dimensionality, making the superiority of the MNMDP evident. Observe again that LDA obtains the better performance than PCA, LPP, and UDP when \( l = 5 \) and less than them when \( l = 6, 7, 8 \). As a consequence, the promising performance of our MNMDP is confirmed on HRF.

| Methods | \( l = 5 \) | \( l = 6 \) | \( l = 7 \) | \( l = 8 \) |
|---------|-------------|-------------|-------------|-------------|
| PCA     | 66.96 ± 1.49 | 71.37 ± 1.27 | 73.87 ± 1.44 | 76.99 ± 2.43 |
| LDA     | 68.64 ± 1.31 | 73.00 ± 1.06 | 75.43 ± 1.29 | 78.55 ± 2.31 |
| LPP     | 67.21 ± 1.41 | 71.58 ± 1.25 | 74.29 ± 1.09 | 77.55 ± 2.01 |
| UDP     | 67.41 ± 1.58 | 71.21 ± 1.26 | 73.36 ± 1.52 | 75.57 ± 2.28 |
| MNMDP   | 74.31 ± 1.62 | 78.47 ± 1.61 | 81.21 ± 1.64 | 83.78 ± 1.58 |

4.2. Experiments on FKP. The PolyU FKP (Finger-Knuckle-Print) database [46] contains 165 subjects, each of which has 48 samples that were taken in two separate sessions. Each session per subject has 6 images for each of the left index finger, the left middle finger, the right index finger, and the right middle finger. Here, we use FKP ROI database obtained by ROI extraction algorithm in [47] and the size of each image is \( 220 \times 110 \) pixels, with 256 grey levels per pixel. To save computation time, we use a subset of the FKP ROI database in the experiment. The data set we selected contains the 100 subjects and 12 images for each individual. Note that the first 3 sample images per finger are selected. In our experiments, each image is resized to 32 × 32 pixels for computational efficiency. As an example, Figure 4 shows twelve images of one subject in the FKP database. We form the training set by a random subset of \( l \) images per class and use the rest as a testing set. In the following experiments on FKP, the numbers of training sample images per class are chosen as \( l = 5, 7, 9, 11 \).

In the experiments on FKP, we first investigate the recognition performance of MNMDP on FKP by varying the \( k \)-neighborhood parameter \( W_k \). The value of \( W_k \) is varied from 1 to 21 in Step 2. The maximal average recognition rate at each value of \( W_k \) for each \( l \) is illustrated in Figure 5. It is obvious that the classification performance of MNMDP increases when the number of training samples increases.
Figure 3: The average recognition rates versus dimension on HRF.

Figure 4: Twelve sample images of one subject in the FKP database.

As can be noted in Figure 5, the performance of MNMDP nearly increases when \( W_k \) changes from 1 to 7 at first and then decreases when \( W_k \) becomes large. From Figure 5, it can be concluded that the appropriate value of \( W_k \) in MNMDP is significant for capturing the geometrical structure and pattern discrimination of data on FKP, the same as on HRF, and it can be easily selected to obtain good performance.

To further verify the performance of MNMDP, the comparative classification results of the competing methods on FKP are reported in Figure 6 with varying the reduced dimensionality from 5 to 100 in Step 5. According to the experimental results in Figure 5, the best \( k \)-neighborhood parameters for MNMDP are set as \( W_k = 7 \) for \( l = 5, 7, 9 \) and \( W_k = 5 \) for 11, respectively. We can see that the classification performance of each method almost ascends monotonically with increasing the dimensionality, at first increases quickly, and finally increases very slowly or even keeps stable. It is noticeable that the proposed MNMDP outperforms the other methods significantly across all values of dimensionality for each \( l \), and UDP almost obtains the worst performance among them. We can also observe that the performance of LDA is always superior to that of LPP and UDP over a large range of dimensionality, and PCA performs better than LPP when the dimension is small whereas it does worse when the dimension is large. Based on the experimental
comparisons in Figure 6, the conclusion we have drawn is that our method consistently obtains better classification performance, irrespective of the variation in dimensions.

For each $l$ on FKP, the experimental comparisons of the competing methods in terms of the maximal average recognition rates with the corresponding standard deviations (stds) and values of dimension in the parentheses are also tabulated in Table 2. It should be noted that the best performance for each method increases as the number of the training samples increases. We can also observe that the best performance of LDA is better than that of PCA, LPP, and UDP and UDP and LPP are preferable to PCA. Through the comparative study of the best performance of the competing methods, we can conclude that the MNMDP has more discriminating power to achieve the satisfactory classification.

| Methods | $l=5$ | $l=7$ | $l=9$ | $l=11$ |
|---------|-------|-------|-------|--------|
| PCA     | 63.31±2.52 | 76.56±2.80 | 85.18±1.98 | 90.75±2.40 |
|         | (110)    | (125)   | (130)   | (110)   |
| LDA     | 65.80±2.35 | 79.44±2.88 | 88.02±2.01 | 92.80±2.07 |
|         | (105)    | (115)   | (125)   | (125)   |
| LPP     | 64.30±2.43 | 78.36±2.52 | 86.80±1.80 | 91.50±2.31 |
|         | (110)    | (110)   | (120)   | (105)   |
| UDP     | 64.88±2.32 | 78.16±2.84 | 86.37±1.83 | 91.30±2.45 |
|         | (110)    | (120)   | (125)   | (135)   |
| MNMDP   | 69.49±1.98 | 82.96±2.54 | 90.98±1.59 | 94.45±1.90 |
|         | (70)     | (50)    | (55)    | (65)    |

In the experiments on AR, the classification performance of the proposed MNMDP versus the $k$-neighborhood parameter $W_k$ is first carried out for each $l$, shown in Figure 8. Notice that the values of $W_k$ are varied from 1 to 21 in Step 2. It is obvious that the performance of MNMDP increases with the increase of the training samples. As can be noted in Figure 8, the recognition rates of MNMDP ascend quickly at first with increase of $W_k$ and then almost keep stable when $W_k$ becomes large. Thus, we can conclude that the appropriate value of $W_k$ plays an important role in MNMDP for preserving the geometry of data and enhancing the power of pattern discrimination, and it can be easily set to obtain good performance.

Moreover, the classification performance of the proposed MNMDP is further evaluated on AR by varying the reduced dimensionality, in comparison with the competing methods. The dimensionality varies from 5 to 100 with an interval of 5. Note that from the results in Figure 8, the best $k$-neighborhood parameters for MNMDP are set as $W_k = 17$ for $l = 7$, $W_k = 15$ for $l = 9$, and $W_k = 13$ for $l = 11, 13$, respectively. The performance of each method in terms of average recognition rates is illustrated in Figure 9. It can be seen that the performance of each method for each $l$ first increases rapidly when $W_k$ becomes large and then approximately tends to be stable. Compared to PCA, LDA, LPP, and UDP, the proposed MNMDP method almost has the best performance by varying the dimensionality, especially at the large values of dimensionality. In the meantime, LDA is superior to PCA, LPP, and UDP with increasing the dimensionality. In addition, in most cases LPP is better than PCA and UDP, and PCA is better than UDP. From the comparative performance in Figure 9, we can conclude that our method always has better classification results over a large range of the dimensionality.

The maximal average recognition rates of each competing method on AR for each $l$ with the corresponding standard deviations (stds) and values of dimension in the parentheses are also reported in Table 3. It is to be noted that the best performance among them is described in bold face. We can see that the performance of each method is improved by increasing the number of the training samples. As Table 3
displays, MNMDP has the best performance among the methods for each \( l \). It can also be observed that the best classification performance of LDA is better than that of LPP, PCA, and UDP. Consequently, the experimental results in Figure 9 and Table 3 on AR face database certainly demonstrate the good performance of the proposed MNMDP.

4.4. Experiments on Musk. The Musk (version1) database [49] is one of the two-class classification tasks that predicts whether new molecules will be musks or nonmusks. It totally contains 476 samples, each of which has 166 attributes that depend on the exact shape or conformation of these molecules. In the experiments, we set the number of training samples per class as \( l = 50, 80, 110, 140 \), respectively, and the remaining samples are used to test the competing methods.

In the experiments on Musk, we first investigate the classification performance of the proposed MNMDP versus the \( k \)-neighborhood parameter \( W_k \) for each \( l \). The experimental results are shown in Figure 10. The values of \( W_k \) are presented from 1 to 21 in Step 2. As can be seen in Figure 10,
The average recognition rates of MNMDP versus $W_k$ on AR.

Table 3: The maximal average recognition rates (%) of each method on AR with the corresponding standard deviations (stds) and values of dimension in the parentheses.

| Methods | $l = 7$ | $l = 9$ | $l = 11$ | $l = 13$ |
|---------|---------|---------|---------|---------|
| PCA     | 64.17 ± 2.05 | 71.60 ± 1.40 | 76.70 ± 3.28 | 81.80 ± 3.46 |
| LDA     | 70.94 ± 1.74 | 78.24 ± 1.09 | 83.80 ± 2.82 | 88.80 ± 2.04 |
| LPP     | 65.23 ± 1.79 | 72.60 ± 1.02 | 78.20 ± 2.67 | 83.10 ± 3.31 |
| UDP     | 65.27 ± 1.68 | 72.22 ± 0.96 | 77.43 ± 3.17 | 82.80 ± 4.18 |
| MNMDP   | 81.63 ± 3.24 | 89.52 ± 2.99 | 94.37 ± 2.26 | 99.00 ± 0.94 |

To further verify the classification performance of our MNMDP on Musk, it is compared to the competing methods by varying the reduced dimensionality. The dimensionality increases from 1 to 30 in Step 1. It should be noted that from the results in Figure 10, the best k-neighborhood parameters for MNMDP are determined as $W_k = 5$ for $l = 50, 80, 110, 140$, respectively. Figure 11 shows the performance of each method in terms of average recognition rates. It can be found that the performance of each method for each $l$ first ascends at small values of $W_k$ and then approximately tends to be stable or increases slowly when $W_k$ becomes large. As shown in Figure 11, the proposed MNMDP method almost has the best performance by varying the dimensionality among all the methods. It can also be observed that PCA, LPP, and UDP get the similar performance when $W_k$ is about larger than 10, and LDA obtains the worse performance when $W_k$ varies from 6 to 30. Therefore, the classification results in Figure 11 indicate that our method is always better than PCA, LDA, LPP, and UDP with the change of dimensionality.

The comparative experiments on Musk for each $l$ in terms of the maximal average recognition rates with the corresponding standard deviations (stds) and values of dimension in the parentheses are finally shown in Table 4. Note that the best performance for each $l$ among all the methods is represented in bold face. It is clear that the classification performance of the proposed MNMDP is better than PCA, LDA, LPP, and UDP. In the meantime, the optimal reduced dimensionality of our MNMDP for each $l$ is smaller than that of them. Therefore, we can conclude that the proposed MNMDP does well in dimensionality reduction with good classification.

In summary, the proposed MNMDP almost yields the best classification performance in all the experiments, compared to PCA, LDA, LPP, and UDP. It implies that both pattern discrimination and geometrical information of the data are very important for classification, and MNMDP fully captures them in the learning processing.

5. Discussions

In this section, some characteristics of the proposed MNMDP that are available for classification are discussed. We first analyze the affinity weight for intraclass and interclass neighborhoods, then discuss the MNMDP from the viewpoint of distance metric learning [50, 51], and finally explore the eigenvalues of the generalized eigenvalue problem in MNMDP.

According to (6) and (13), the affinity weight can be thought of as integration of the local weight $\exp(-\|x_i - x_j\|^2/\delta)$ and the discriminating weight that is divided into two categories: intraclass one for intraclass neighbors $1 + \exp(-\|x_i - x_j\|^2/\delta)$ and interclass one for interclass neighbors...
$1 - \exp(-\|x_i - x_j\|^2/\delta)$. It means that the affinity weight cannot only preserve local structures of data but also distinguish between different classes. Through the analysis of the affinity weight in MNMDP, its three properties can be summarized as follows [23].

Property 1. For a given point, the affinity weight gives more similarity to the intraclass neighbors than the interclass ones, when the Euclidean distances are equivalent. This is favorable for classification.

Property 2. The affinity weight can retain intraclass and interclass similarity in certain ranges no matter how strong the noise is, since $1 \leq 1 + \exp(-\|x_i - x_j\|^2/\delta) \leq 2$ and $0 \leq 1 - \exp(-\|x_i - x_j\|^2/\delta) \leq 1$. This can largely preserve local geometric structures of data for different classes.

Property 3. The affinity weight can strengthen the ability of margin augmentation and noise suppression. This is explained by two aspects. One is that the close points from different classes could have smaller values of similarity and the margin between different classes can be augmented, because $1 - \exp(-\|x_i - x_j\|^2/\delta)$ tends towards 0 with a decrease of the Euclidean distance. The other is that the affinity weight can inhibit the noise in some degree; that is, the more distant points from the same class could be less similar to each other, because $\exp(-\|x_i - x_j\|^2/\delta)$ tends towards 0 with an increase of the Euclidean distance.

The above good properties make a good supervised construction of the affinity graph, and they are well maintained in the process of MNMDP subspace learning. They can overcome the issue that the interclass neighbors might have
the same representation as the intraclass neighbors in the reduced subspace for the given point in LPP [52]. Based on them, MNMDP achieves a good projection that carries not only local geometric structures but also discriminant information. Thus, the properties of the affinity weight make the MNMDP robust and effective for the classification tasks.

In regard to classification, the MNMDP can be viewed as one method of distance metric learning [50, 51]. Once the transformation matrix $\Phi$ is obtained by MNMDP, the squared distance between $x_i$ and $x_j$ in the embedded subspace can be formulated as follows:

$$d(y_i, y_j) = \|\Phi^T x_i - \Phi^T x_j\|^2$$
$$= (x_i - x_j)^T \Phi \Phi^T (x_i - x_j)$$
$$= (x_i - x_j)^T M (x_i - x_j)$$
$$= \|x_i - x_j\|^2_{M'}$$

(26)

where $M = \Phi \Phi^T$. Thus, finding a liner transformation $\Phi$ in MNMDP is equivalent to learning the distance metric $M$ implicitly. This is very important in pattern classification. Moreover, making the projection orthogonal (i.e., $\Phi^T \Phi = I$) will tend to preserve distances between any points $x_i$ and $x_j$ in the projected subspace [53]. Due to linearity, the overall geometry of the data will also tend to be preserved. It should be noted that there are many distance metric learning methods that can be used for dimensionality reduction, such as neighborhood components analysis (NCA) [54] and large margin nearest neighbor classification (LMNNC) [55]. Since NCA and LMNNC closely connect to the $k$-NN classifier and their objective functions and optimization solutions are very different from MNMDP, more comparative discussions between them are beyond the scope of this paper here.

Given that data points are projected along an eigenvector $\phi_i$ corresponding to eigenvalue $\lambda_i$, it follows from (24) with the orthogonal constraint $\phi_i^T \phi_i = 1$ that

$$J(\phi_i) = J^-(\phi_i) - J^+(\phi_i)$$
$$= \phi_i^T (S^- - S^+) \phi_i$$
$$= \phi_i^T \lambda_i \phi_i$$
$$= \lambda_i.$$  

(27)

It is clear that $J(\phi_i)$ characterizes the margin between interclass and intraclass neighborhoods for all the points along the projection $\phi_i$ in terms of $\lambda_i$, and the margin here can be regarded as a measure for the misclassification degree. If $\lambda_i > 0$, then $J^-(\phi_i) > J^+(\phi_i)$, and samples may tend to be correctly classified. The larger the value of $\lambda_i$ (s.t. $\lambda_i > 0$) is, the easier the classification is. In practice, we can only select $d$ leading eigenvectors to form the transformation matrix $\Phi$ corresponding to $d$ dominant positive eigenvalues and omit all the eigenvectors with relatively small eigenvalues. In this way, a low-dimensional subspace can be obtained. As reported above in our experiments, MNMDP really achieves the best classification with low dimensionality. If $\lambda_i \leq 0$, then $J^-(\phi_i) < J^+(\phi_i)$, and samples may be put into wrong classes. As a consequence, we discard those eigenvectors in respect to nonpositive eigenvalues. It should be noted that the classification performance is seriously degraded when all eigenvalues are nonpositive. In this case, it is difficult to distinguish interclass and intraclass points along the projections. To solve the problem, we can map data points into Hilbert space with some kernel tricks. However, this is out of the scope of this paper and will be discussed in the future work.

Based on the discussions aforementioned, the proposed MNMDP has more discrimination for classification. From the perspective of the classification, the key point is to
enhance the pattern discrimination between samples from different classes of data. For the high-dimensional data, the good way is to find a projection that makes the samples from the same class compact and ones from different classes separable in the low-dimensional space. To visually verify the superiority of MNMDP for classification, we project the sample images in HRF database (described in Section 4.1) onto a two-dimensional subspace with the competing methods, and the samples from the first three subjects of HRF in the projected space with each method are illustrated in Figure 12. Note that, since the samples of each subject are collected in two separated sessions with different variations, each subject is clustered into two subclasses with each method. We can obviously observe that the separability of class clusters of MNMDP is much better than that of PCA, LDA, LPP, and UDP and the subclasses in MNMDP become more compact. Moreover, there is no overlap between three classes in MNMDP, compared to the other methods. Figure 12 intuitively validates the good discriminating capability of MNMDP for classification. The results are consistent with the observation from the experiments in Section 4 and the analysis in this section.

6. Conclusions

In this paper, the MNMDP introduced is a liner supervised dimensionality reduction technique, which can well preserve the local geometric structures of data and fully use class information for classification. In the proposed MNMDP, we employ distinct affinity weight for both intraclass and interclass neighbors of all points and then keep the margin between intraclass and interclass neighborhoods maximum.
Figure 12: Data distribution of each method in two-dimensional projection space, where the three classes are denoted by “+”, “O”, and “◻”, respectively.
through a liner mapping. In order to well investigate the classification performance of the MNMDP, our experiments are conducted on four high-dimensional databases, in comparison with the competing methods: LPP, UDP, PCA, and LDA. Through the comprehensive experiments, it demonstrates the effectiveness and robustness of the proposed MNMDP with satisfactory performance in pattern classification. In the future work, we plan to extend the MNMDP with some kernel tricks.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the National Science Foundation of China (Grant nos. 6170126, 61202045, 61005017, 61202110, and 61300228). The authors wish to thank the anonymous reviewers for their valuable suggestions.

References

[1] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: a review," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 4–37, 2000.

[2] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 586–591, June 1991.

[3] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 711–720, 1997.

[4] A. M. Martinez and A. C. Kak, "PCA versus LDA," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 2, pp. 228–233, 2001.

[5] X. He and P. Niyogi, "Locality preserving projections," in Proceedings of the Advances in Neural Information Processing Systems, 2003.

[6] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using laplacianfaces," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 3, pp. 328–340, 2005.

[7] M. H. Yang, "Kernel eigenfaces vs. kernel fisherfaces: face recognition using kernel methods," in Proceedings of the IEEE 5th International Conference on Automatic Face Gesture Recognition, pp. 215–220, 2002.

[8] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a Kernel Eigenvalue problem," Neural Computation, vol. 10, no. 5, pp. 1299–1319, 1998.

[9] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," Neural Computation, vol. 12, no. 10, pp. 2385–2404, 2000.

[10] J. Yang, A. F. Frangi, J.-Y. Yang, D. Zhang, and Z. Jin, "KPCA plus LDA: a complete kernel fisher discriminant framework for feature extraction and recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 2, pp. 230–244, 2005.

[11] G. Feng, D. Hu, D. Zhang, and Z. Zhou, "An alternative formulation of kernel LPP with application to image recognition,

[12] J.-B. Li, J.-S. Pan, and S.-C. Chu, "Kernel class-wise locality preserving projection," Information Sciences, vol. 178, no. 7, pp. 1825–1835, 2008.

[13] J. Yang, D. Zhang, A. F. Frangi, and J.-Y. Yang, "Two-dimensional PCA: a new approach to appearance-based face representation and recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 1, pp. 131–137, 2004.

[14] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "MPCA: multilinear principal component analysis of tensor objects," IEEE Transactions on Neural Networks, vol. 19, no. 1, pp. 18–39, 2008.

[15] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," in Proceedings of the Advances in Neural Information Processing Systems, 2004.

[16] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H.-J. Zhang, "Multilinear discriminant analysis for face recognition," IEEE Transactions on Image Processing, vol. 16, no. 1, pp. 212–220, 2007.

[17] X. He, D. Cai, and P. Niyogi, "Tensor subspace analysis," in Proceedings of the Advances in Neural Information Processing Systems, 2005.

[18] D. Hu, G. Feng, and Z. Zhou, "Two-dimensional locality preserving projections (2DLPP) with its application to palmprint recognition," Pattern Recognition, vol. 40, no. 1, pp. 339–342, 2007.

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

[20] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," Neural Computation, vol. 15, no. 6, pp. 1373–1396, 2003.

[21] J. Yang, D. Zhang, J.-Y. Yang, and B. Niu, "Globally maximizing, locally minimizing: unsupervised discriminant projection with applications to face and palm biometrics," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 4, pp. 650–664, 2007.

[22] H. Wang, S. Chen, Z. Hu, and W. Zheng, "Locality-preserved maximum information projection," IEEE Transactions on Neural Networks, vol. 19, no. 4, pp. 571–585, 2008.

[23] H. Zhao, S. Sun, Z. Jing, and J. Yang, "Local structure based supervised feature extraction," Pattern Recognition, vol. 39, no. 8, pp. 1546–1550, 2006.

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

[25] W. Yu, X. Teng, and C. Liu, "Face recognition using discriminant locality preserving projections," Image and Vision Computing, vol. 24, no. 3, pp. 239–248, 2006.

[26] W. K. Wong and H. T. Zhao, "Supervised optimal locality preserving projection," Pattern Recognition, vol. 45, no. 1, pp. 186–197, 2012.

[27] D. Zhong, I. Han, X. Zhang, and Y. Liu, "Neighborhood discriminant embedding in face recognition," Optical Engineering, vol. 49, no. 7, Article ID 077203, pp. 1–7, 2010.

[28] J. Gou and Z. Yi, "Locality-based discriminant neighborhood embedding," The Computer Journal, vol. 56, no. 9, pp. 1063–1082, 2013.
X. Fan, Y. Xu, X. Li, Z. Fan, H. Liu, and Y. Chen, "Locality preserving discriminant projections for face and palmprint recognition," Neurocomputing, vol. 73, no. 13-15, pp. 2696–2707, 2010.

D. Cai, X. He, J. Han, and H.-J. Zhang, "Orthogonal laplacian faces for face recognition," IEEE Transactions on Image Processing, vol. 15, no. 11, pp. 331–341, 2010.

G.-F. Lu, Z. Jin, and J. Zou, "Face recognition using discriminant sparsity neighborhood preserving embedding," Knowledge-Based Systems, vol. 31, pp. 19–27, 2012.

F. Yin, L. C. Jiao, F. Shang, S. Wang, and B. Hou, "Fast fisher sparsity preserving projections," Neural Computing and Applications, vol. 23, pp. 691–705, 2013.

Y. Chen, Z. Li, and Z. Jin, "Feature extraction based on maximum nearest subspace margin criterion," Neuro Processing Letters, vol. 37, no. 3, pp. 355–375, 2013.

X. Fang, Y. Xua, X. Li, Z. Fan, H. Liu, and Y. Chen, "Locality and similarity preserving embedding for feature selection," Neurocomputing, 2013.

J. Yang, D. Chu, L. Zhang, Y. Xu, and J. Yang, "Sparse representation classifier steered discriminative projection with applications to face recognition," IEEE Transactions on Neural Networks and Learning Systems, vol. 24, no. 7, pp. 1023–1035, 2013.

J. Wang and I. Hua, "Supervised discriminant projection with its application to face recognition," Neuro Processing Letters, vol. 34, no. 1, pp. 1–12, 2011.

H. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17, no. 1, pp. 1157–1165, 2006.

"The PolyU HRF database," http://www4.comp.polyu.edu.hk/~biometrics/HRF/HRF_old.htm.

X. Wu, V. Kumar, J. R. Quinlan et al., "Top 10 algorithms in data mining," Knowledge and Information Systems, vol. 14, pp. 1–35, 2008.

"The PolyU FKP database," http://www.comp.polyu.edu.hk/~biometrics/FKP.htm.

L. Zhang, L. Zhang, D. Zhang, and H. Zhu, "Ensemble of local and global information for fingerknuckle-print recognition," Pattern Recognition, vol. 44, no. 9, pp. 1990–1998, 2011.