Weighted Locally Linear Embedding Algorithm for Classification

Shan-Wen Zhang[1,2] De-Shuang Huang[2]

[1,2]*Department of Engineering and Technology, Xijing University, Xi'an 710123*

[2]*Hefei Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, Anhui, 230031, China.*

Keywords:
LLE
Weighted LLE
Robust PCA (RPCA)
Outlier

A B S T R A C T

Locally linear embedding (LLE) has been shown to be effective in discovering the geometrical structure of the data. But when it is applied to real-world data, it shows some weak points, such as being quite sensitive to noise points and outliers, being unsupervised in nature and failing to discover the discriminant structure in the data. To address these problems, an improved version of LLE, namely weighted LLE algorithm, is presented in this paper. When constructing the nearest-neighbor graph, the reliability values of *k* neighbor points are computed based on weighted PCA and Iterative Re-weighted Least Square (IRLS), which indicate how likely the neighbor points are clean data points. Then a geodesic distance measurement is designed by taking three aspects into account, including the reliability values of their *k* neighbor points, their class information and their local information. The proposed method using the distance measurement can improve the dimension reduction and internal feature extraction performance. The experiments on synthetic data and real FERET data demonstrate that the suggested algorithm can efficiently maintain an accurate low-dimensional representation of the noisy manifold data with less distortion, and acquire higher average recognition rates of FERET compared to others.

1. Introduction

Manifold learning is an important dimensionality reduction tool which can discover the intrinsic structure of high dimensional data and provide understanding of multidimensional patterns in data mining, pattern recognition and machine learning. Manifold learning algorithms can be applied to extract the intrinsic features of observed data in high dimensional space by preserving the local geometric characteristics. However, due to the locality geometry preservation, most existing manifold learning methods, such as locally linear embedding (LLE) [1], are not robust against outliers in the data and are in general sensitive to noise and outliers. LLE may lose its efficiency in dimension reduction for classification since it is built based on Euclidean distance for constituting the neighborhood information, and since the local Euclidean distance does not match the classification property generally. That is to say, two sample points belonging to different classes may also have a short Euclidean distance. This phenomenon may result in that the neighbors of a point may come from different classes. Often they fail in the presence of the high dimensional outliers or noise, because the outliers or noise may distort the local structure of the manifold. When the local outliers or noise level are increased, the mapping efficiency quickly become very poor. In the process of initial data mapping, if the influence of outliers or noise have not been well eliminated or suppressed, the topological structure of initial data will not be well kept in low dimensional space, which is extremely essential to manifold learning. Recently, some improvement measures have been presented for improving the robustness of noisy manifold learning. Hadid et al. [2] proposed an efficient LLE. This method is based on the assumption that all outliers are very far away from the data on the manifold. But this assumption is not always satisfactory for many real-world data. Based on the weighted PCA, Zhang et al. [3] proposed a preprocessing method for removing outliers and suppressing noise before ordinary manifold learning is performed. However, the method for determining the weights is

heuristic in nature without formal justification. Similar to the method of Zhang et al. [3], Park et al. [4] proposed a method for outlier handling and noise suppression by using weighted local linear smoothing for a set of noisy points sampled from a nonlinear manifold. Chang et al. [5] proposed a robust LLE (RLLE) based on the robust PCA, which is very robust against outliers. But RLLE would also fail when the sample density is low or the data points are unevenly sampled, or the data have small perturbation. Hein et al. [6] proposed a denoising method based on a graph-based diffusion process of the point sample. This method has good performance for noisy data, but when the outliers are closer to another data component, wrong transformation may happen. Zhang et al. [7] proposed a modified LLE, which can solve the above problems of LLE. But it would fail in the case as the data points are not distributed on or close to a two-dimension or three-dimension nonlinear manifold. Yin et al. [8] proposed a neighbor smoothing embedding (NSE) for noisy data, which can efficiently maintain an accurate low-dimensional representation of the noisy data with less distortion. But this method introduces an additional parameter and ignores the statistical feature. Pan et al. [9] proposed a weighted LLE (WLLE), which can optimize the process of discovering the intrinsic structure by avoiding unreasonable neighbor searching and is robust parameter changes. But in WLLE, the weighted distance is not always the geodesic distance in real-data.

Above all modified LLE algorithms are unsupervised in nature. They do not utilize the class information of the data. Recently, in order to address this problem, many supervised LLE (SLLE) algorithms have been boomed using the sample label information [10-13]. Unlike LLE, SLLE projects high dimensional data to the embedded space using class membership relations which can generate well-separated clusters in the embedded space. SLLE introduces an additional tuning parameter α, controlling how much of the class label information should be taken into account. It can be effectively applied to the classification tasks.

The purpose of this paper is to improve the robustness and classification ability of LLE by designing a new distance measurement. The main contributions of the paper are as follows:

(1) Based on robust PCA and IRLS, the weighted values of neighbor points are estimated, which can reflect the reliability of the data points;

(2) A weighted distance measurement is proposed by using the reliability values of the neighbor points, the class information and the local information of the data points;

(3) The weighted LLE algorithm is presented based on the weighted distance measurement.

The proposed method is tested on synthetic data and real FERET data with artificial noise added, the results of experiments demonstrate that the method not only is more robust to noisy data, but also has better performance of classification.

The remainder of this paper is organized as follows. Section 2 briefly introduces the noise influence upon manifold learning. Section 3 discusses weighted PCA and robust PCA, and proposes a weighted distance measurement. Section 4 presents a weighted LLE. Section 5 demonstrates the method with synthetic data and compares the experimental results of our method on real FERET data with other manifold learning methods. Finally, some conclusive remarks and future works are proposed in Section 6.

2. Noise influence upon manifold learning

As to noise influence upon manifold learning, it is already described in some correlated literatures [8,9,14,15]. The following are two actual examples indicating that noise affect manifold dimensionality reduction. These data belong to artificial data. Fig.1 shows how the classical LLE works in finding the low dimensional embedding of the S-curve and Swiss-roll manifold from $\mathbb{R}^3$ to $\mathbb{R}^2$.

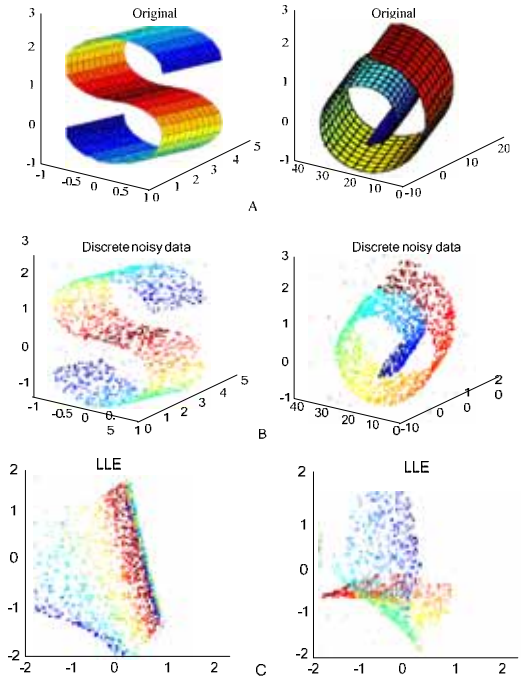


Fig.1. LLE applied to the S-curve and Swiss-roll dataset, respectively. (A) S-curve and Swiss-roll manifold in $\mathbb{R}^3$; (B)Discrete points randomly sampled from the manifold (totally 1,500 clean points from S-curve and 150 noisy points added; totally 1,800 clean points from Swiss-roll and 200 noisy

points added); (C) LLE result embedding space in $\square^2$ with the nearest neighbors k=15 based on Euclidean distance.

As we can see clearly from Fig.1(C), the LLE algorithm is topologically unstable; it cannot preserve well the local geometry of the data manifolds in the embedding space in the presence of noise, the entire data topological structure is also destroyed. By further experiments, we are able to find that small errors in the data connectivity (topology) can lead to large errors in the solution. Choosing a very small neighborhood is not a satisfactory solution, as this can detach the manifold into a large number of disconnected regions. In the presence of outliers (noise points), the *K*-nearest neighbors of a neat point on the manifold may no longer lie on a locally linear patch of the manifold, leading to a small bias to the reconstruction. As for an outlier (noise point), its neighborhood is typically much larger than that of a clean point.

Through observing the embedding results of two above noisy data, we can see that noise influence upon manifold data mainly reflects the destruction of topological structure; such destruction means the failure of manifold algorithm, therefore the key problem of whether manifold learning can be used successfully is how to validly deal with noise or outliers in initial data.

3. Weighted distance measurement

Suppose that there are n data points in initial data space, denoted as X=[X_1, X_2,…, X_n], $X_i \in R^D$, which includes clean and noisy points. Using the algorithm of Weighted PCA (WPCA), Robust PCA (RPCA) [5,16] and Iterative Reweighted Least Square (IRLS) [17], all neighbor points are weighted. Based on the weighted values of the neighbor points and its class information, a weighted distance measurement is presented.

3.1 Weighted PCA

PCA is frequently adopted for dimensionality reduction as the mean square error upon reconstruction is minimized. However, when the data is nonlinear and from multi-modes, performing PCA is far from satisfactory and thus its local extensions, i.e., WPCA and RPCA, have been suggested.

In manifold learning, there usually is a local linear hypothesis, i.e., any data point X_i together with its k nearest-neighbor points $X_{i1}, X_{i2}, ..., X_{ik}$ is located on a linear super-plane, so X_i and $X_{i1}, X_{i2}, ..., X_{ik}$ are situated on a local linear plane of manifold. So for these points some linear transform methods can be used to map the mentioned points onto d-dimensional subspace:

$$Z_j = B^T \left(X_{ij} - V \right) \in R^d \tag{1}$$

where Z_j is the mapped results of X_{ij} in low dimensionality space, V is the translation vector; $B=[b_1, b_2, ..., b_d] \in R^{D \times d}$ is a linear transform matrix, which satisfying orthogonal normalized condition:

$$b_j^T b_i = \delta_{ji}, 1 \le j, i \le k \tag{2}$$

Similarly, it is possible to inverse-map Z_j onto a high dimensional space:

$$X_{ij}^{'} = V + BB^T \left(X_{ij} - V \right) = V + BZ_j \tag{3}$$

where $X_{ij}^{'}$ is the image of X_{ij} by inverse-mapping.

Theoretically, by mapped and inverse-mapped, a point and its image are the same, yet in actual application there are still errors, i.e.

$$\varepsilon_j = X_{ij} - X_{ij}^{'} = X_{ij} - V - BB^T \left(X_{ij} - V \right) \tag{4}$$

So the WPCA can be employed to estimate linear transformation matrix and translation vector, given a set of nonnegative weighted vector $A_i = [a_1, a_2, ..., a_k]$ for the k neighbor points, the optimization problem becomes minimizing the total weighted square error:

$$\begin{aligned} E_{WPCA} &= min\{\sum_{j=1}^{k} a_j \left\| \varepsilon_j \right\|^2\} \\ &= min\{\sum_{j=1}^{k} a_j \left\| X - VI^T - BZ \right\|^2\} \end{aligned} \tag{5}$$

where a_j is a weighted value corresponding to the error ε_j, I is a vector of all elements is 1.

By adopting Laplacian number multiplication, it is possible to find the least mean square estimation (LMSE) of linear translation vector:

$$V = \sum_{j=1}^{k} a_j X_{ij} \Big/ \sum_{j=1}^{k} a_j \tag{6}$$

The LMSE of linear transformation matrix can be obtained through finding the solution of feature decomposition:

$$Sb = \lambda b \tag{7}$$

where S can be expressed as following:

$$S = \frac{1}{k} \sum_{j=1}^{k} a_j (X_{ij} - V) \left(X_{ij} - V \right)^T \tag{8}$$

Eq. (7) indicates that LMSE of linear transformation matrix is composed of the eigenvector corresponding to least eigenvalue of matrix S.

To reduce the influence of possible noisy points or outliers among the k neighbor points, we would like to set A_i such that noisy points or outliers get small weight values. In other words,

if a neighbor point x_{ij} has a large error norm $\|\varepsilon_j\|$, we would like to set aj small. Robust estimation methods can help to set the appropriate weights, making weighted PCA a robust version of PCA.

3.2 Robust PCA

In many signal processing problems, observed signals are rather complex-valued and /or are distorted by outliers or spiky (impulsive) noise. For processing such data it is necessary to use the RPCA which is capable of resisting outliers (large errors).

Through Eq. (8), it can be seen that S refers to weighted covariance matrix of k nearest-neighbor samples. When Laplacian number multiplication is used to find the solution of linear transformation matrix and translation vector, the weighted vector $a=[a_1,a_2,...,a_k]$ is usually assumed to be fixed. However, it is only the ideal case. In actual applications, a_j is variable, and also there is a function relation between the a_j and the error ε_j. Refer to Huber's robust estimation [18], the function relation between weighted value and error is set to satisfy the following:

$$\psi(\varepsilon)=\begin{cases}1 & |\varepsilon|\le\mu\\ \exp(-\dfrac{(\varepsilon-\mu)^2}{2\sigma^2}) & otherwise\end{cases} \tag{9}$$

where μ,σ are mean and variance respectively, defined as $\mu=1/k\sum_{i=1}^{k}\varepsilon_i$ and $\sigma^2=1/k\sum_{i=1}^{k}(\varepsilon_i-\mu)^2$.

From Eq.(9), it is known that the solution of weighted value ψ can be found by the error ε whose solution is found through linear transformation B and translation vector V, meanwhile B and V are dependent upon ψ, i.e., through ε, there exist an independent relation between ψ and ε. In light of their cyclic iterative relation, the solution of optimal weighted value ψ can be obtained by IRLS algorithm [17], which is described as following:

*Step*1 Initializing: utilize standard PCA on the neighborhood of X_i to find the solution of linear transformation matrix B and translation vector V, denoted as $B^{(0)}$ and $V^{(0)}$, and set t=0;

*Step*2 Iterative process:

(1) $t=t+1$;

(2) Calculate error

$$\varepsilon_j^{(t-1)}=X_{ij}-V^{(t-1)}-B^{(t-1)}[B^{(t-1)}]^T(X_{ij}-V^{(i-1)}),1\le j\le k;$$

(3) Count weighted value

$$a_j^{(t-1)}=\psi(\varepsilon_j^{(t-1)}),1\le j\le k;$$

(4) Compute the least weighted estimation $B^{(t)}$ and $V^{(t)}$ according to the Formula (6)-(8);

(5) Until the $B^{(t)}$ and $V^{(t)}$ do not change too much from the last-obtained $B^{(t-1)}$ and $V^{(t-1)}$.

*Step*3 Execute the above-said iteration processes for each point so as to acquire a weighted vector $a=[a_1,a_2,\ldots,a_k]$.

3.3 Weighted geodesic distance

Let $\{X_i,C_i\}_{i=1}^{n}\subset R^n\times\{1,2,...,m\}$ be a set of n data points which belong to m different classes, where $X_i\in R^D$, $D\ll n$ and C_i is ith sample label, $1\le C_i\le m$. For any data point X_i and its k neighborhood points $X_{i1},X_{i2},\ldots,X_{ik}$, after WPCA and IRLS, each neighbor point X_{ij} has its weighted value a_j, the normalized weighted value a_j^* of X_{ij} is computed as $a_j^*=a_j/\sum_{u=1}^{k}a_u$, which can be regard as reliability measure for each neighbor. This normalized weighted value a_j^* can serve as a reliability measure for each neighbor point X_{ji}. For all points not in the neighborhood of X_i, their weighted values are set to zero.

After performing WPCA and IRLS for all points X=[X_1, X_2,…, X_n], a total reliability value of X_i, denoted as ρ_i, is obtained by summing up the a_i^* from all WPCA and IRLS runs. The smaller the value ρ_i for the point X_i, the more likely it is that X_i is a noisy point or outlier.

Based on above analysis, a novel distance measurement is defined as follows:

$$d(X_i,X_j)=\|X_i-X_j\|+\alpha\cdot M\cdot(1-S(X_i,X_j)) \tag{10}$$

where $M=\max_{i,j}\|X_i-X_j\|$ is the data diameter in Euclidean distance, $\alpha\in[0,1]$ is a tuning parameter which controls the amount to which class information should be incorporated, $S(X_i,X_j)$ is expressed as

$$S(X_i,X_j)=\begin{cases}1, & \text{If } C_i=C_j\\ \sqrt{\rho_i\cdot\rho_j}, & \text{If } C_i\ne C_j \text{ and, } X_i\in\Delta_{kj} \text{ or } X_j\in\Delta_{ki}\\ 0, & \text{If } X_i\notin\Delta_{kj} \text{ and } X_j\notin\Delta_{ki}\end{cases} \tag{11}$$

where ρ_i,ρ_j is the normalized weighted values of neighbor points X_i, X_j, respectively; Δ_{ki}, Δ_{kj} is the k nearest neighbor of X_i, X_j respectively.

From Eq.(10) and Eq.(11), with the class information and the reliability of neighbor point being added to weighted distance $d(X_i,X_j)$, the properties of the $d(X_i,X_j)$ can be summarized as follows:

(1) When the Euclidean distance is equal, the inter-class $d(X_i,X_j)$ is larger than the intra-class $d(X_i,X_j)$, which reflects the class information of data points. That is to say, two points with smaller distance most likely have the same label. On the contrary, the distance between two points with different labels will be larger. This benefits to sample classification;

(2) The smaller the ρ_i or ρ_j is, the larger the $d(X_i,X_j)$ is then the more likely the neighbor point is nosy point.

(3) With the increasing of the Euclidean distance, the intra-class $d(X_i,X_j)$ is equal to the Euclidean distance, while the inter-class $d(X_i,X_j)$ increases faster.

(4) This distance $d(X_i,X_j)$ with certain ability to "recognize" noise in the data. On the one hand, the intra-class distance is usually small. So the larger the distance $d(X_i,X_j)$ is, the more possibly the noise exists. On the other hand, the inter-class distance $d(X_i,X_j)$ is usually large. So the smaller it is, the more possibly the noise exists. Both aspects indicate that $d(X_i,X_j)$ can gradually strengthen the power of noise suppression.

4. Weighted LLE algorithm

To preserve the integrity of the data, all data points including the clean data points and the outliers are projected into the embedding space. Embedding is achieved by using the weighted geodesic distance to define the neighborhood. With the weighted distance measurement, our weighted LLE algorithm is presented as follows:

*Step*1: For each data point X_i, determine the set Δ_i of k nearest neighbors of X_i by K-nearest-neighbor algorithm in Euclidean space. In this step, standard Euclidean metric is used to select the nearest neighbors.

*Step*2: Compute WPCA, RPCA and IRLS on the set Δ_i and obtain the weighted values of the neighbor points, then normalize and sum up the weighted values to gain the reliability of each data point, then construct the weighted distance measurement, as shown in Eq.(10).

Step 3: For each data point X_i, re-determine the set Λ_i of k nearest neighbors by K-nearest-neighbor algorithm in the weighted distance measurement space according to Eq.(10).

Step 4: Compute the local combination weights W_{ij} that best reconstruct, i.e. which minimize the constrained least square problem:

$$\min \left\| X_i - \sum_{X_j \in \Lambda_i} X_j W_{ij} \right\| \tag{12}$$

Intuitively, W_{ij} reveals the layout of the points around X_i, it subject to two constraints, $\sum_{X_j \in \Delta_i} W_{ij} = 1$ and $W_{ij} = 0$ for any $X_j \notin \Delta_i$. In the first constraint, each point is represented as a convex combination of its neighbor; the second reflects that weighted LLE is a *local* method.

Step 5: Solve the optimization problem for the low-dimensional embedding $Y=[Y_1,Y_2,\ldots,Y_n]$ which best preserves the local geometry represented by the reconstruction weights. This means to solve,

$$\min \sum_{i=1}^{n} \left\| Y_i - \sum_{X_j \in \Lambda_i} Y_j W_{ij} \right\|^2 \tag{13}$$

where Y subject to two constraints $\sum_{i=1}^{n} Y_i = \boldsymbol{0}$ and $\sum_{i=1}^{n} Y_i Y_i^T = n \cdot I$ (normalized unit covariance), $\boldsymbol{0}$ is a column vector of zeros and I is an identity matrix. Two constraints are appended to remove the translational degree of freedom and the rotational degree of freedom, respectively.

Based on the weighted matrix W, Y can be obtained by the eigenvectors of the sparse, symmetric and positive semi-definite matrix M as follows: $M=(I-W)(I-W)^T$, where $W=\{W_{ij}\}$ corresponding to the second to $(d+1)$th smallest eigenvalues, d is the dimension of the vectors Y_i.

Note that Eq.(13) can be expressed in a quadratic form $\sigma(Y) = \sum_{ij} M_{ij} Y_i Y_j^T$, where $M = [M_{ij}]_{n \times n}$. By the Rayleigh-Ritz theorem [19], minimizing Eq. (13) can be done by finding the eigenvectors with the smallest (nonzero) eigenvalues of the sparse matrix M.

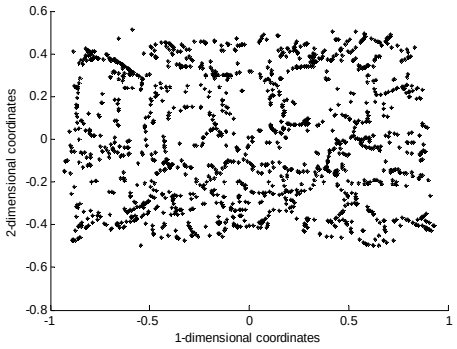
5. Experiments and results

In the experimental section, we explore the effect of proposed alogorithm on two noisy manifold data: S-curve and FERET. The Gaussian noise is added to these data. In the processing, we map the noisy data points onto low dimensional space by our weighted LLE. In classification of the FERET experiment, we choose K-NN as the classifier. In proposed method there are a number of parameters which need to be decided, among which the neighbor k is the most important.

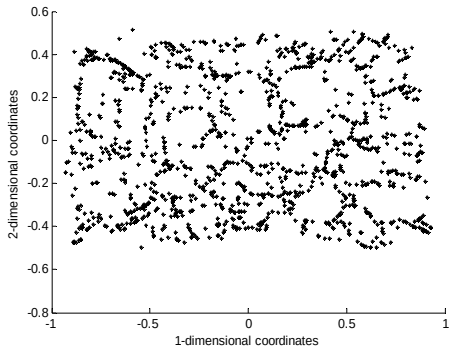
5.1 Noisy data of S-curve

The first experiment is done on S-curve. We take 1500 points arbitrarily from S-curve. These points are unlabeled. That is to say, for any two points X_i and X_j, $C_i \neq C_j$, where C_i, C_j are the

labels of X_i and X_j, respectively. Then we adopt classical LLE and our weighed LLE to study the mapping data, the neighbor size k is set to 15, the tuning parameter α is set to 0.5. Fig.2 shows the 2-D visualized results using LLE and weighed LLE. It can be clearly seen from Fig.2 that for non-noisy data, the mapping results by LLE and weighed LLE are almost the same.



(A) Mapping result by LLE



(B) Mapping result by weighed LLE

Fig.2. Mapping result on S-curve by LLE and weighted LLE

In order to examine the robustness of our weighed LLE, 150 noise points are added to the known 1500 points. These noise points are at least at a certain distance from the 1500 initial data points on the manifold. Fig.3 is mapping result by weighted LLE. From Fig.3, it can be observed that the noise points among the mixed data can be validly separated by our suggested algorithm. So the 2-D visualized results by weighted LLE can keep quite well topological structure of the data.

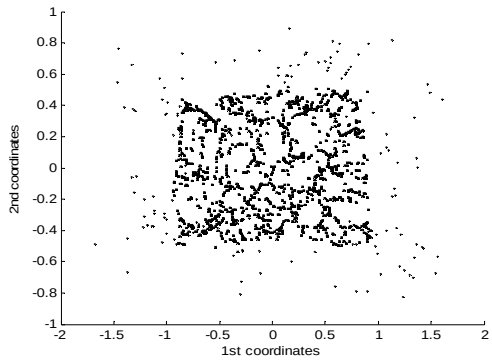


Fig.3. Mapping result by weighed LLE

Through the above 2-D visualized results, it is verified that our method is by and large feasible.

5.2 FERET recognition

Face recognition is one of the most popular research topics in pattern recognition. It can be widely used in authentication, entertainment, information security, and intelligent robotics. Recently, great development has been done by many researchers on both algorithm and systems. A key question in face recognition is the dimension reduction for feature extraction. In this section, the performance of our weighted LLE is evaluated on FERET datasets and compared with the performances of RLLE [5], WLLE [9] and SLLE [11]. The FERET dataset, available at http://www.cs.toronto.edu/~roweis/data.html [20], is becoming a standard database for testing and evaluating feature extraction methods for face recognition.

In this experiment, a subset is selected from the original FERET dataset [21]. It contains 200 individuals and seven images for each person. It is composed of images whose names are marked with two-character strings: “ba” “bd” “be” “bf” “bg” “bj” and “bk”. This subset involves two facial expression images, two left pose images, two right pose images and an illumination image. In all the experiments, preprocessing is performed to crop the original images. For face data, the original images were normalized such that the two eyes were aligned at the same position, then the facial areas were cropped into the final images for matching. The size of each cropped image in the first experiment is 80×80 pixels and 32×32 pixels in the second experiment with 256 gray levels per pixel. The sample images of one person are shown in Fig.4, where (A) is original images, (B) and (C) are the Gaussian noisy images with variance=0.1 and variance=0.4, respectively. It can be seen from Fig.4C that the face images are seriously distorted by noise. After RLLE, WLLE, SLLE and our proposed algorithm have been applied to extracting classification features, the 1-NN classifier is adopted to predict the labels of the data for its simplicity.

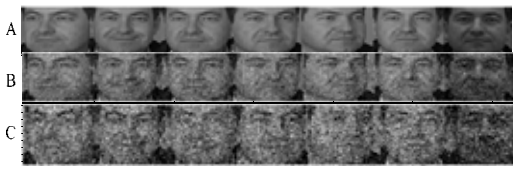


Fig.4. Sample images of one object of the FERET database

The training set and test set are divided in the same way as in [21]. Firstly, 4 images are selected randomly as training set and the rest 3 images as test set, without overlapping between the

two sets. We built the classification models using the training samples, and estimated the classification correct rates using the test set. In this paper, 20-fold cross-validation tests are performed to evaluate the performance of RLLE, WLLE, SLLE and our method. All numerical experiments are performed with 20 random splitting of the original dataset. The number of nearest neighborhood is set to k=l-1=4-1=3, where l denotes the number of training samples per class. The justification for this parameter is that each sample should connect with the remaining l-1 samples of the same class provided that within-class samples are well clustered in the observation space [22]. The tuning parameter α is set to 0.5 for SLLE and our proposed method, the threshold parameter is set to 0.5 for RLLE. After features have been extracted by performing RLLE, WLLE, SLLE and our method, the original face images are project to a subspace at 120 dimensions, the 1-NN Classifier is adopted to predict the labels of the test data. In each experiment, we select the optimized recognition results at corresponding dimensions for different feature extraction methods. Table.1 shows the mean and standard deviation (in parenthesis) of accuracies from 20 experiments for each method with the dimensionality. Since the random splits for training and test set are disjoint, the results given in Table.1 should be unbiased. Fig.5 shows the recognition rate versus the variation of dimension, where the variance of the noise is 0.1. From Table.1 and Fig.5, it can be found that the recognition result of our proposed method is best than other three methods.

Table 1 The recognition rates of FERET images

Method / FERET	RLLE [5]	WLLE [9]	SLLE [11]	Weighted LLE (Ours)
Original data	80.65 (1.20)	79.78 (1.50)	81.34 (1.23)	81.25 (1.39)
Noisy(Var=0.1)	72.58 (4.46)	73.48 (3.24)	70.18 (4.23)	74.34 (4.02)
Noisy(Var=0.4)	69.29 (5.25)	70.66 (4.07)	66.59 (4.82)	73.63 (4.45)

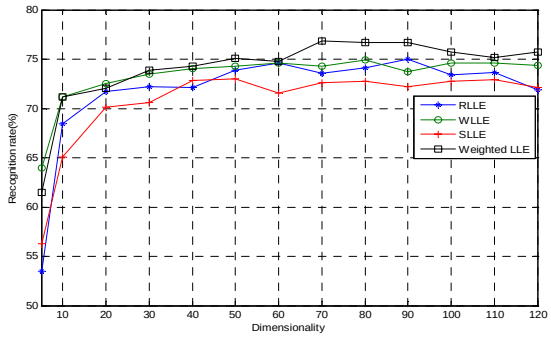


Fig.5. Recognition rate of RLLE, WLLE, SLLE and weighted LLE versus the dimension on noisy FERET data (variance=0.1).

Because the key parameter for LLE-based methods is the number of neighbor k, we record the optimal recognition rates with different k, where the tuning parameter α is set to 0.7; the threshold is set 0.5 for RLLE. Fig.6 shows the optimal recognition rates with the varied neighbor k on noisy FERET data, where the variance is 0.1.

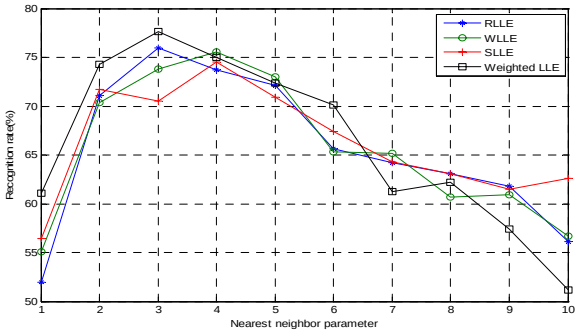


Fig.6 Best recognition rates of RLLE, WLLE, SLLE and weighted LLE with the varied k on noisy FERET data (variance=0.4).

In the following experiment, we set k=3 for SLLE and set k=4 for weighted LLE, the tuning parameter α is set from 0.1 to 1 with step 0.1 for SLLE and Weighted LLE. The performances of SLLE and Weighted LLE with varied α is displayed in Fig.7.

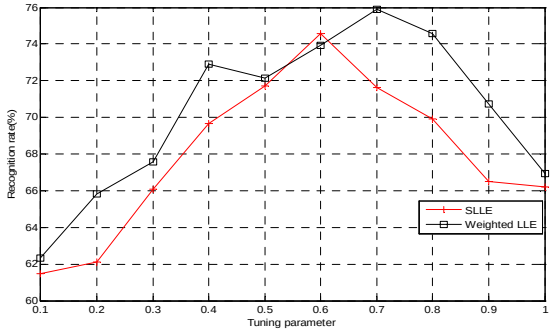


Fig.7 Best recognition rates of SLLE and weighted LLE with the varied α on noisy FERET data (variance=0.1).

From above facial recognition experiments, we can obtain the better parameters for our proposed method, i.e., d=70, k=3, α =0.7, and in this case, the recognition rate is 75.89% on the data which variance is 0.1. Of course, it's only the primary conclusion from our limited experiments. In the future more experiments should be done to verify the conclusion.

5.3 Discussion

Several experiments on the real FERET data have been systematically performed. These experiments have revealed some interesting points:

1) All four methods performed better in the original FERET data. Moreover, the SLLE is a little better than other three methods. But the results on the noisy data indicate the SLLE is not robust again the noise.

2) When adding the noise to data, our proposed method outperformed the other methods. This is probably due to the

proposed weighted distance measurement.

3) From Fig.7, we find that the tuning parameter α is important to the supervised LLE and our proposed method.

6. Conclusions

Manifold learning has aroused a great deal of interests in dimensionality reduction, but most of manifold learning methods are sensitive to outliers and noise, and are unsupervised. To address these issues, a weighed LLE algorithm is presented to develop the robust and the classification ability of traditional LLE. In the process of algorithm, RPCA and IRLS algorithm is first adopted to give weighted values to every neighbor point. Then a geodesic distance measurement between two data points of the graph are designed by taking three aspects into account, including the weighted values of their k neighbor points, their class information and their local information. The experiment results show that the proposed method is effective and robust to noise. However, our method still has some shortcomings. Its computational requirement is significantly higher than that of LLE. The main problem lies in the computation of the weighted values of neighbor points by RPCA and IRLS. Since the RPCA and IRLS procedure have to be executed for each data point, multiple iterations are usually needed. Our future work is how to bring the computational complexity of mapping new samples down further and how to speed up the RPCA procedure. It would be interesting to compare the proposed method to other combinations of nonlinear mapping methods and classifiers on a number of data sets, to gain insight into what methods are suitable for what class of problems.

It should be declared that the geodesic distance measurement proposed in this paper for weighted LLE may be extended to make other nonlinear dimension reduction methods, such as supervised Isomap, more robust. It is a potential direction for future research.

Acknowledgements

This work was supported by the grants of the National Science Foundation of China, Nos. 60975005, the grant from the National Basic Research Program of China (973 Program), No.2007CB311002.

Reference

1. Roweis S T, Saul L K, 2000. Nonlinear dimensionality reduction by locally linear embedding. Science, 290 (5500): 2323-2326.
2. Hadid and M. PietikÄainen,2003. Efficient locally linear embeddings of imperfect manifolds. In Proceedings of the Third International Conference on Machine Learning and Data Mining in Pattern Recognition, 188-201, Leipzig, Germany, 5-7.
3. Zhang, Z., Zha, H., 2003. Local linear smoothing for nonlinear manifold learning. CSE-03-003, Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA, USA.
4. Park, J., Zhang, Z., Zha, H., Kasturi, R., 2004. Local smoothing for manifold learning. CVPR'04 2, 452-459.
5. Chang, H., Yeung, D.Y., 2006. Robust locally linear embedding. Pattern Recognition 39 (6), 1053–1065.
6. Hein M., Maier, M., 2006. Manifold denoising. Advances in NIPS 20, Cambridge, MA:561-568.
7. Z.Y. Zhang, J. Wang, MLLE: Modified locally linear embedding using multiple weights, in: B. Scholkopf, J.C. Platt, T. Hoffman (Eds.), Advances in Neural Information Processing Systems, vol. 19, MIT Press, Cambridge, MA, 2007, pp. 171-181
8. Junsong Yin, Dewen Hua, Zongtan Zhou. Noisy manifold learning using neighborhood smoothing embedding. Pattern Recognition Letters 29 (2008) 1613-1620
9. Yaozhang Pan, Shuzhi Sam G, Abdullah Al Mamun. Weighted locally linear embedding for dimension reduction. Pattern Recognition 42 (2009) 798- 811
10. D. De Ridder, O. Kouropteva, O. Okun, M. Pietikainen, R.P.W. Duin, Supervised locally linear embedding, in: Proc. Joint Int. Conf. ICANN/ICONIP 2003, New York, in: Lecture Notes in Computer Science, vol. 2714, 333-341.
11. Pillati M. and Viroli C.," Supervised Locally Linear Embedding for Classification: An Application to Gene Expression Data Analysis," In: Proceedings of 29th Annual Conference of the of the German Classification Society, pp.15-18,2005.
12. Dong Liang, Jie Yang, Zhonglong Zheng, Yuchou Chang. A facial expression recognition system based on supervised locally linear embedding. Pattern Recognition Letters,Vol.26,2374-2389,2005.
13. Kouropteva Olga, Okun Oleg, Pietikäinen Matti. Supervised locally linear embedding algorithm for pattern recognition. Pattern recognition and image analysis, LNCS, vol. 2652, 386-394,2003.
14. Balasubramanian M., Schwartz, E. L.,2002. The LLE algorithm and topological stability. Science, 295(5552):7.
15. Brand, M., Charting a manifold. Advances in Neural Information Processing Systems 15, 15,2003.
16. Mia Hubert and Sanne Engelen. Robust PCA and classification in biosciences. Bioinformatics,Volume 20, Number 11, 1728-1736
17. Oleary, D. P., 1990. Robust Regression Computation Using Iteratively Reweighted Least-Squares. Siam Journal on Matrix Analysis and Applications, 11(3):466-480.
18. P.J. Huber. Robust regression: asymptotics, conjectures, and Monte Carlo. Annals of Statistics, 1(5):799-821, 1973.
19. Horn, R.A. and Johnson, C.R. Matrix Analysis. Cambridge University Press, Cambridge,UK, (1990)
20. P.J. Phillips, H. Moon, S.A. Rizvi, and P.J. Rauss, 2002.The FERET Evaluation Methodology for Face-Recognition Algorithms, IEEE Trans. Pattern Analysis and Machine Intelligence, 22(10), 1090-1104. Yale Univ. Face Database,http://cvc.yale.edu/projects/yalefaces/yalefaces.html.
21. Bo Li, De-Shuang Huang, ChaoWang, Kun-Hong Liu, Feature extraction using constrained maximum variance mapping. Pattern Recognition 41 (2008) 3287-3294
22. Jian Yang, David Zhang, Jing-yu Yang, and Ben 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, APRIL 2007