2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Similarity-balanced Discriminant Neighborhood Embedding

Chuntao Ding
School of Computer Science and Technology & Provincial Key Laboratory for Computer Information Processing
Suzhou, China
Email：20124227036@suda.edu.cn

Li Zhang
School of Computer Science and Technology Provincial Key Laboratory for Computer Information Processing
Suzhou, China
Email：zhangliml@suda.edu.cn

Yaping Lu
School of Computer Science and Technology Provincial Key Laboratory for Computer Information Processing
Suzhou, China
Email：20134227010@stu.suda.edu.cn

Shuping He
School of Computer Science and Technology Provincial Key Laboratory for Computer Information Processing
Suzhou, China
Email：sphe@suda.edu.cn

***Abstract*—The idea that with the help of proper dimensionality reduction, trying to make the samples with the same label be compact and the ones with the different labels be separate after projection, is introduced into classification problems with high-dimensional data. Based on the analysis of the drawbacks of Discriminant Neighborhood Embedding (DNE) and Locality-Based Discriminant Neighborhood Embedding (LDNE), being the two relatively successful Locally Discriminant Analysis methods proposed in recent years, this paper proposes a method called Similarity-balanced Discriminant Neighborhood Embedding (SBDNE). When constructing the adjacent graph, SBDNE fully takes into account the geometric construction of manifold and the problem of imbalance between the intra-class points and the inter-class points. By endowing these two kinds of samples with different similarities and selecting the near neighbors according to the similarity matrix, not only the structure in the original space can be preserved more efficiently, but also the choice of discriminative information increases. The method proposed here has a better recognition with comparisons to some classical methods, which fully shows that SBDNE method has the capacity to efficiently solve the classification problem.**

***Keywords*—discriminant neighborhood embedding; adjacent graph; intra-class; inter-class**

I. INTRODUCTION

There are more and more study and application fields that need to deal with high-dimensional data. As a result, to achieve its analysis and visualization, we have to reduce the dimensionality so as to make the high-dimensional data embed into relatively low-dimensional feature subspace with the inner structure of data preserved. This skill is widely used in the fields such as computer vision, machine learning and pattern recognition and so on.

As classical methods, both Principal Component Analysis (PCA) [1-2] and Linear Discriminant Analysis (LDA) [2-3] assume that the data processed are from the Euclidean space. However, the manifold learning [4-6], rising after 2000, shows that many complex objects are situated in some manifold subspace and their non-linear inner structure cannot be learned via traditional methods. Nevertheless, manifold learning algorithms only consider the training samples and cannot get an explicit mapping, so they cannot perform incremental learning for new data, namely called the out-of-sample problem, because of which manifold learning methods are under restrictions for classification. To cover this shortage, He et al. proposed Locality Preserving Projection (LPP) [7] and Neighborhood Preserving Embedding (NPE) [8], both of which can directly map the new sample into a low-dimensional subspace via the projection matrix obtained by the training procedure.

Dimensionality reduction methods are composed of unsupervised ones and supervised ones. The former focuses on the better representations of high-dimensional data without considering the labels, and the latter tries to achieve the classification efficiently with the labels employed. They are also called represented dimensionality reduction and discriminative dimensionality reduction. As a classical linear discriminative dimensionality reduction method, LDA tries to find a projection direction, being conducive to discriminate, by minimizing the divergence of samples with the same class and maximizing the divergence of samples with the different classes. Although LDA has been widely used in pattern recognition field, it still has the problem of the small sample size and requires the data to obey a Gaussian distribution. However, the practical data often dissatisfy the hypothesis, so, to overcome this drawback, maximum margin criterion (MMC) [18] and margin Fisher analysis (MFA) [9-10] methods have been proposed. MMC is mainly to maximize the difference between inter-class and intra-class scatters and MFA as an extension of LDA, MFA is able to efficiently solve the problems discussed above. For MFA, the locally structure of samples is preserved by constructing the homogeneous and heterogeneous neighbor adjacency graphs, and the optimal projection direction is found by minimizing the ratio of the sum of distance between the samples with the same class and the sum of distance between the samples with the different classes.

Dimensionality reduction methods can also be divided into the non-graph-structure-based ones and the graph-structure-based ones. The former directly reduces the dimensionality without taking into account the structure information of data in the original space, and the latter makes the geometric structure

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

of data in high-dimensional space still be preserved in low-dimensional space by constructing the preserved structure graph. As a relatively classical graph-structure-based method, LPP achieves to preserve the local structure of original data by making the samples being close to each other in original space and still be close to each other in low-dimensional space. However, LPP is an unsupervised method that does not efficiently utilize the label information. , which might degrade their performance in pattern recognition. Based on the idea of LPP, many methods have been proposed, such as Supervised Locality Preserving Projections (SLPP) [11] and Neighborhood Discriminant Projection for Face Recognition (NDP) [12] and so on, we can easily see that these supervised methods mainly make use of class label information to well guide the procedure of dimensionality reduction. Among which the Discriminant Neighborhood Embedding (DNE) [13] proposed by Zhang et al. is a much efficient method. For DNE method, first, by constructing an adjacent graph, the relationship between the samples in original space and their neighbors, including the same class and the different classes, is preserved, then make the samples have the same structure in the low-dimensional space, and finally by employing the spectral analysis the dimensionality of discriminative subspace is calculated. However, DNE cannot preserve the detailed position relationship between the samples and their neighbors, including the same class and the different classes. As a consequence, the recognition rate in low-dimensional space would decrease when the data are unbalanced. By constructing a different adjacent graph with DNE method and endowing different weights, Locality-Based Discriminant Neighborhood Embedding (LDNE) proposed in [14] makes the optimization problem change to optimize the difference between the distance of samples with the same class and the distance of samples with the different classes.

Based on the analysis of the drawbacks of DNE and LDNE, this paper proposes a new supervised dimensionality reduction method called Similarity-balanced Discriminant Neighborhood Embedding (SBDNE). By introducing a new similarity function, SBDNE endows the data with the same class and the different class with different similarity functions. The similar neighbors are selected based on the matrix composed of the similarity functions. By constructing the structure graphs of the samples with the same class and the samples with the different classes and utilizing the geometric structure of manifold, the unbalanced problem between the same class and the different classes is solved. As a result, not only the structure in the original space can be preserved more efficiently, but also the choice of discriminative information increases. Finally, experimental results on the artificial dataset, ORL face dataset , Yale face dataset and FERET face dataset show the effectiveness of SBDNE.

II. RELATED WORK

In this section, we review both DNE and LDNE, which are supervised dimensionality reduction methods. Suppose we have the training samples $\{\mathbf{x}_i, y_i\}_{i=1}^N, \mathbf{x}_i \in R^d, y_i = \{1,2,...,c\}$, where y_i is the label of $\mathbf{x}_i$. c, N and d respectively denotes the number of classes, the number of samples and the dimensionality of samples. The purpose of DNE and LDNE is to find a linear projection that maps the data in the d-dimensional space into the r-dimensional subspace, such as $\mathbf{v}_i = \mathbf{P}^T\mathbf{x}_i$ where $\mathbf{v}_i$ represents the low-dimensional data after projection and $\mathbf{P} \in R^{d\times r}$ is the projection matrix.

A. Discriminant Neighborhood Embedding

DNE aims to make the samples with the same label form a compact sub-manifold and the distance between the samples with the different labels are as far as possible in the low-dimensional subspace after projection. The process of DNE method is as follows:

(1) Define an adjacent graph $\mathbf{F}$, of which the element F_{ij} is given by

$$F_{ij} = \begin{cases} +1 \ , & \mathbf{x}_i \in \aleph_k(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \aleph_k(\mathbf{x}_i) \text{ and } (y_i = y_j) \\ \text{-}1 \ , & \mathbf{x}_i \in \aleph_k(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \aleph_k(\mathbf{x}_i) \text{ and } (y_i \neq y_j) \\ 0 \ , & \text{otherwise} \end{cases} \qquad (1)$$

where $\aleph_k(\mathbf{x}_j)$ denotes the set of k nearest neighbors of $\mathbf{x}_j$, y_i and y_j respectively represent the labels of $\mathbf{x}_i$ and $\mathbf{x}_j$.

(2) Feature mapping: Optimize the following objective function:

$$\begin{cases} \min \ \sum_{ij} \| \mathbf{P}^T\mathbf{x}_i - \mathbf{P}^T\mathbf{x}_j \|^2 \ F_{ij} \\ s.t. \quad \mathbf{P}^T\mathbf{P} = \mathbf{I}. \end{cases} \qquad (2)$$

where $\mathbf{I}$ is the identity matrix, and $\mathbf{P}$ is the projection matrix. Through a simple derivation, the optimization problem changes to be as follows:

$$\begin{cases} \min_{\mathbf{P}} \ tr(\mathbf{P}^T\mathbf{XLX}^T\mathbf{P}) \\ s.t. \qquad \mathbf{P}^T\mathbf{P} = \mathbf{I} \end{cases} \qquad (3)$$

where $\mathbf{L} = \mathbf{D} - \mathbf{F}$, $\mathbf{D}$ is a diagonal matrix with $D_{ii} = \sum_j F_{ij}$ and $tr(\cdot)$ is the trace of matrix. Finally, the projection matrix $\mathbf{P}$ can be obtained by the decomposition of the proper value of a matrix according to the following objective function:

$$\mathbf{XLX}^T\mathbf{P} = \lambda\mathbf{P} \qquad (4)$$

where the optimal projection matrix $\mathbf{P}$ is composed of the r eigenvectors corresponding to the r minimum eigenvalues.

B. Locality-Based Discriminant Neighborhood Embedding

Based on DNE and by endowing the adjacent graph with different weights, LDNE is able to preserve the nearest neighbors. Moreover, it also tries to find an optimal projection matrix by maximizing the difference between the aggregation of samples with the same class and the divergence of samples with the different classes. The process of LDNE method is as follows:

(1) According to the k nearest neighbors rule, construct a similarity matrix $\mathbf{S}$ by

$$S_{ij}=\begin{cases} -\exp\left(\dfrac{-\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^2}{\beta}\right), & if \quad (y_i=y_j) \text{ and } \\ & (\mathbf{x}_i\in\aleph_k(\mathbf{x}_j) \text{ or } \mathbf{x}_j\in\aleph_k(\mathbf{x}_i)) \\ +\exp\left(\dfrac{-\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^2}{\beta}\right), & if \quad (y_i\neq y_j) \text{ and } \\ & (\mathbf{x}_i\in\aleph_k(\mathbf{x}_j) \text{ or } \mathbf{x}_j\in\aleph_k(\mathbf{x}_i)) \\ 0 \quad , & \text{otherwise} \end{cases} \quad (5)$$

where $\beta>0$ is the parameter selected by users.

(2) Feature mapping: Optimize the following objective function:

$$\begin{cases} \max \sum_{ij} \| \mathbf{P}^T\mathbf{x}_i-\mathbf{P}^T\mathbf{x}_j \|^2 S_{ij} \\ s.t. \quad \mathbf{P}^T\mathbf{P}=\mathbf{I}. \end{cases} \quad (6)$$

Through a simple derivation, the optimization problem changes to be as follows:

$$\begin{cases} \max_{\mathbf{P}} \; tr(\mathbf{P}^T\mathbf{XHX}^T\mathbf{P}) \\ s.t. \qquad \mathbf{P}^T\mathbf{P}=\mathbf{I} \end{cases} \quad (7)$$

where $\mathbf{H}=\mathbf{D}-\mathbf{S}$, $\mathbf{D}$ is a diagonal matrix, of which the diagonal elements are composed of the sum of $\mathbf{S}$ by row or by column, such as $D_{ii}=\sum_j S_{ij}$.

Being similar to the DNE method, the projection matrix $\mathbf{P}$ of LDNE can also be obtained by the decomposition of the proper value of a matrix according to the following objective function:

$$\mathbf{XHX}^T\mathbf{P}=\lambda\mathbf{P} \quad (8)$$

where $\mathbf{P}$ is composed of the r eigenvectors corresponding to the r maximum eigenvalues.

C. The Drawbacks of DNE and LDNE

According to the discussion above, we know that when constructing an adjacent graph, DNE only endows the samples with the same label with +1 and the ones with the different labels with -1, which would lead to three drawbacks. First, the locally structure information of data cannot be preserved. Second, sometime, it cannot be efficiently act on the samples with the same label and the ones with the different labels at the same time. Third, when the data are unbalanced, all the nearest neighbors of an example may completely belong to the same class or the different classes so that when constructing an adjacent graph it cannot find the association between the samples with the same label or the different labels. As a consequence, DNE may not find the most efficient sub-manifold.

For LDNE method, it achieves to preserve the locally structure information of data by calculating the similarity between the example and its neighbors. But it has two drawbacks. On the one hand, it is not obvious to distinguish the relationship between the same class and the different classes since they are endowed with the same similarity function. On the other hand, being similar to DNE, when the data are unbalanced, it may not find the most efficient sub-manifold as well.

III. SIMILARITY-BALANCED DISCRIMINANT NEIGHBORHOOD EMBEDDING

To overcome the drawbacks of DNE and LDEN, this paper proposes a new supervised sub-manifold learning algorithm called Similarity-balanced Discriminant Neighborhood Embedding (SBDNE). In detail, it is able to efficiently make the samples with the same label be aggregated and the ones with the different classes be separated in the low-dimensional subspace so as to get a better classification performance.

A. Similarity function

Suppose we have the training samples $\{\mathbf{x}_i,y_i\}_{i=1}^N$. Then, we define a new similarity function between $\mathbf{x}_i$ and $\mathbf{x}_j$ as follows:

$$G(\mathbf{x}_i,\mathbf{x}_j)=\begin{cases} \exp\left(\dfrac{-\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^2}{\beta}\right)\exp\left(\exp\left(\dfrac{-\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^2}{\beta}\right)+1\right), & \text{if } y_i=y_j \\ \exp\left(\dfrac{-\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^2}{\beta}\right)\exp\left(1-\exp\left(\dfrac{-\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^2}{\beta}\right)\right), & \text{if } y_i\neq y_j \end{cases} \quad (9)$$

From (9), we know that the similarity functions for the samples with the same label and the ones with the different labels are different. Specifically, the previous ones are endowed with larger weights and the latter ones are endowed with smaller weights. Fig. 1 shows the curves of similarity function $G(\mathbf{x}_i,\mathbf{x}_j)$ vs the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. When these two samples belong to the same class, the similarity rapidly decreases with the increase of their distance. If they belong to different classes, the similarity slowly decreases with the increase of their distance. Note that the curve for $y_i=y_j$ always lies above that of $y_i\neq y_j$. Moreover, the similarity degree in the same class situates between 0 and e^2 , but in the different classes the interval changes to be between 0 and 1 so that the similarity degrees for the different classes can be inhibited.

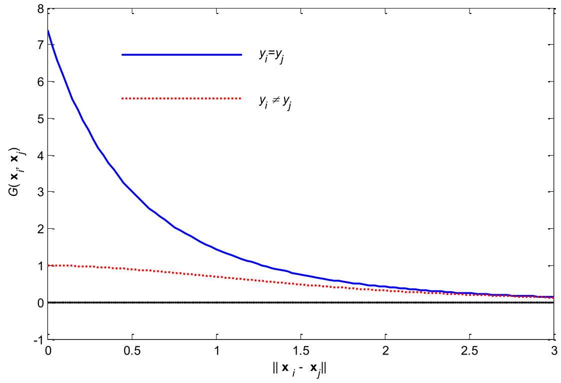
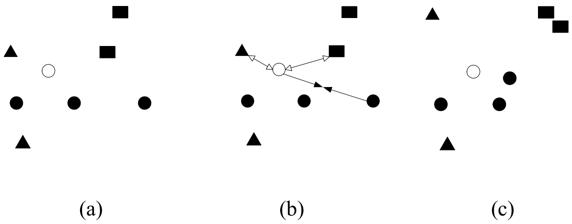


Fig. 1 The similarity between intra-class and inter-class

B. Construction of adjacent graphs

Now, we consider the construction of adjacent graphs according to the new similarity function (9). Our scheme is to select the farthest homogeneous neighbors for a sample to construct an intra-class graph $\mathbf{F}^w$, and it's nearest heterogeneous neighbors to build an inter-class graph $\mathbf{F}^b$. The reason can be illustrated by Fig. 2. In Fig. 2(a), there are three classes denoted by solid square, circle and solid triangle. For the hollow circle point, we select the farthest neighbor in the solid circle points, and the nearest neighbors in the solid square and triangle points as shown in Fig. 2(b). Fig. 2(c) ideally gives their images in the subspace. We expect that the farthest homogeneous could be attracted to around the sample and the nearest heterogeneous neighbors could be pushed way from the sample.



(a) (b) (c)

Fig.2 An illustration. (a) The hollow circle point has seven neighbors. (b) The interactions by attraction and repulsion for the points. (c) Projected points in the subspace.

For F_{ij}^{w}, we select the k homogeneous samples with the smallest similarity for $\mathbf{x}_i$ and preserve their structural relationships. Namely,

$$F_{ij}^{w}=\begin{cases} G\left(\mathbf{x}_i,\mathbf{x}_j\right)\ , & \text{if } \mathbf{x}_i\in N_k^{+}(\mathbf{x}_j)\ or\ \mathbf{x}_j\in N_k^{+}(\mathbf{x}_i) \\ 0\quad , & \text{otherwise} \end{cases} \tag{10}$$

where $N_k^{+}(\mathbf{x}_i)$ and $N_k^{+}(\mathbf{x}_j)$ respectively denote the set of farthest homogeneous neighbors of $\mathbf{x}_i$ and $\mathbf{x}_j$, and $\mathbf{x}_i$ has the same label with $\mathbf{x}_j$. The intra-class compactness has the form:

$$\Phi(\mathbf{P})=\sum_{i,j}\|\mathbf{P}^T\mathbf{x}_i-\mathbf{P}^T\mathbf{x}_j\|^2\ F_{ij}^{w} \tag{11}$$

On the contrary, for F_{ij}^{b}, k heterogeneous nearest neighbors with the highest similarity are selected for $\mathbf{x}_i$. Then,

$$F_{ij}^{b}=\begin{cases} G\left(\mathbf{x}_i,\mathbf{x}_j\right)\ , & \text{if } \mathbf{x}_i\in N_k^{-}(\mathbf{x}_j)\ or\ \mathbf{x}_j\in N_k^{-}(\mathbf{x}_i) \\ 0\quad , & \text{otherwise} \end{cases} \tag{12}$$

where $N_k^{-}(\mathbf{x}_i)$ and $N_k^{-}(\mathbf{x}_j)$ respectively denote the set of nearest heterogeneous neighbors of $\mathbf{x}_i$ and $\mathbf{x}_j$, and $\mathbf{x}_i$ has the different label with $\mathbf{x}_j$. Thus, we have the inter-class divergence as

$$\Omega(\mathbf{P})=\sum_{i,j}\|\mathbf{P}^T\mathbf{x}_i-\mathbf{P}^T\mathbf{x}_j\|^2\ F_{ij}^{b}\quad . \tag{13}$$

By respectively building intra-class structure graph and inter-class structure graph, each example is able to get the associations with the samples with the same or different classes. In other words, for an example, we can get at least two associations, namely the association with the same class and the association with the different classes.

We try to maximize the difference between the nearest inter-class distance and the farthest intra-class distance so as to make the distance between the same classes is nearer and the distance between the different classes is farther in the projection sub-space. That's to say, we need to maximize

$$\Psi(\mathbf{P})=\Phi(\mathbf{P})-\Omega(\mathbf{P})\quad . \tag{14}$$

Through a simple derivation (see Appendix A), the optimization problem changes to be:

$$\begin{cases} \max\limits_{\mathbf{P}}\ \ tr(\mathbf{P}^T\mathbf{X}\mathbf{U}\mathbf{X}^T\mathbf{P}) \\ s.t.\qquad \mathbf{P}^T\mathbf{P}=\mathbf{I} \end{cases} \tag{15}$$

where $\mathbf{U}=\mathbf{D}^b-\mathbf{F}^b-\mathbf{D}^w+\mathbf{F}^w$, $\mathbf{D}^b$ and $\mathbf{D}^w$ are diagonal matrixes with $D_{ii}^b=\sum_j F_{ij}^b$ and $D_{ii}^w=\sum_j F_{ij}^w$, respectively. The detail of this algorithm is shown in Algorithm 1.

```
Algorithm 1 Similarity-balanced Discriminant Neighborhood Embedding (SBDNE)

Input: Training set $\{(\mathbf{x}_i,y_i)\}_{i=1}^N$ ; sample matrix $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N]\in R^{d\times N}$
Output: Projection matrix $\mathbf{P}$
  1). Build the intra-class adjacent graph $\mathbf{F}^w$ and inter-class adjacent graph $\mathbf{F}^b$ according to (10) and (12), respectively.
  2). Perform eigendecomposition on the matrix $\mathbf{XUX}^T$ . Suppose we obtain the eigenvalue $\lambda_i$ and the corresponding eigenvector $\mathbf{p}_i$ , and eigenvalues are organized by descending order, $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_d$ .
  3). Get the $r$ eigenvectors corresponding to the first $r$ eigenvalues, and then we have the projection matrix as $\mathbf{P}=[\mathbf{p}_1,\mathbf{p}_2,\ldots,\mathbf{p}_r]$ .
```

IV. Experiments

In this section, we will discuss applications of SBDNE, and its comparisons with MFA, DNE and LDNE. Both the number of neighbors k for the four methods and β for the SBDNE, and LDNE are the tunable parameters. In our experiments, we select nearest neighbor classifier to classify our data after dimensionality reduction.

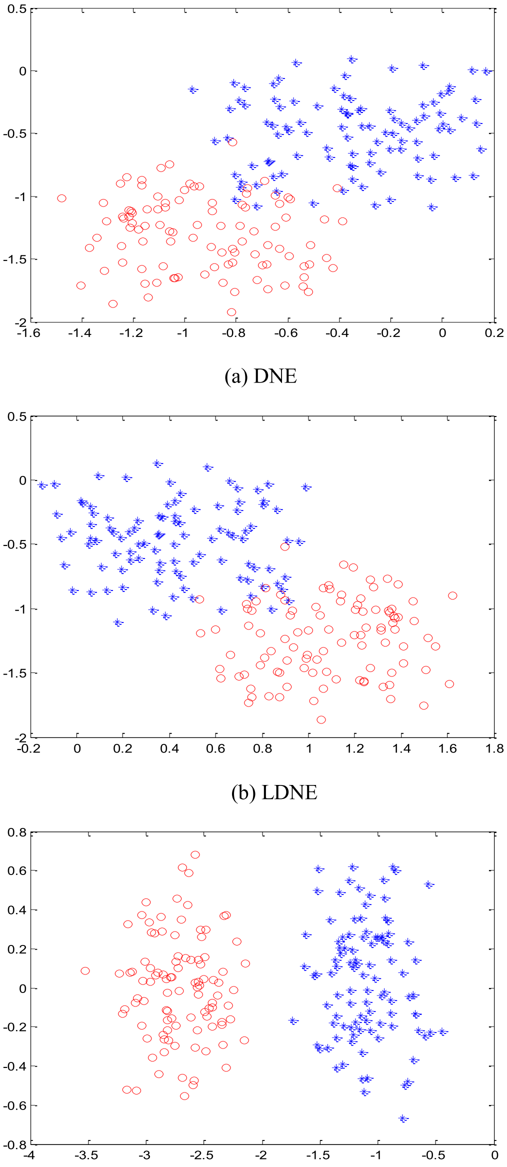
A. Synthetic Dataset

We generate two class samples obeying uniform distribution, ones of which are the random numbers drawn from the interval $\left[0,1\right]^5$, and the other ones are drawn from $\left[0.7,1.7\right]^5$. There are 200 training and 200 test samples. The projection matrix is learned by DNE, LDNE and SBDNE respectively.

In this experiment, k is selected to be 1 and β is got via 10-fold cross-validation for LDNE and SBDNE. The range of β is from 1 to 50. Fig. 3 shows the projected data obtained by DNE, LDNE and SBDNE, respectively.

From Fig. 3, we can know that compared with DNE and LDNE, SBDNE works better for classification since it achieves to make the intra-class samples be aggregated and the inter-class ones be separated. From another point of view, this also indicates that the projection matrix learned by SBDNE is more satisfied to classify the samples. For DNE method, it builds the adjacent graph by exploiting the relationships between the samples and their neighbors without considering the locally position information of the samples. For LDNE method, on the one hand, it is not obvious to distinguish the samples when the intra-class samples and the inter-class ones are endowed with the same similarity function, and on the other hand, being similar to DNE method, it may not find the most efficient sub-manifold. On the contrary, SBDNE method fully takes into account not only the position information of data but also the balanced relationships between the intra-class data and the inter-class data so that it has the better recognition effect.

To verify this, Table I provides the quantitative analysis. Table I presents intra-class scatter, inter-class scatter and the ratio between them, where the intra-class scatter is the sum of distances of all two samples in the same class, the inter-class scatter is the sum of all two samples in the different class. Of course, we expect that the inter-class scatter is large, the intra-class scatter is small, and the ratio of inter-class scatter to intra-class scatter is large. The larger the ratio is, the better the separability is. For the raw data, the ratio is 2.072. The projected data obtained by the three methods has a higher ratio. The inter-class scatters in four cases are almost the same. But SBDNE generates a rather smaller intra-class scatter. Thus, the separability on the projected data obtained by SBDNE is the best.



(a) DNE

(b) LDNE

(c) SBDNE

Fig.3 Projected data obtained by DNE (a), LDNE (b) and SBDNE (c).

TABLE I. SEPARABILITY ON SYNTHETIC DATASET

	Intra-class scatter	**Inter-class scatter**	**Ratio of inter-class scatter to Intra-class scatter**
Raw Data	1.7395 × 104	3.6044 × 104	2.0720
DNE	1.0279 × 104	2.1449 × 104	2.0867
LDNE	1.0252 × 104	2.1901 × 104	2.1362
SBDNE	9.8788 × 103	3.2958 × 104	3.3362

B. *Experiments on Face Recognition*

This experiment is based on the three famous datasets, ORL dataset [16], Yale dataset [17] and FERET subset dataset. For SBDNE, MFA, DNE and LDNE, their performance is measured by recognition rate. In the experiment, the parameters k and β of SBDNE are selected to be several different sets of values so as to observe their effect on the recognition rate. The whole training set is divided into 60% training set and 40% validation set, and the value of parameter β is selected based on the training result on the validation set. Finally, all the methods employ the Nearest Neighbors as their classifier.

C. *ORL Face DataSet*

The ORL face dataset [16] consists of 400 face images of 40 persons, with 10 images for each person. Some images are taken at different times so that the person's face expression and face detail may have the different degrees of variation such as open eyes or closed eyes, simile or not simile and with glasses or without glasses. Additionally, face posture changes with deep or plane rotation to 20 degree and face size also has the 10% variation. Each image has the grayscale from 0 to 255 with digitization and normalization and is scaled to be 32x32 (this means an image has 1024 features) for the efficient computation. Fig. 4 shows the images of one person from the ORL dataset.

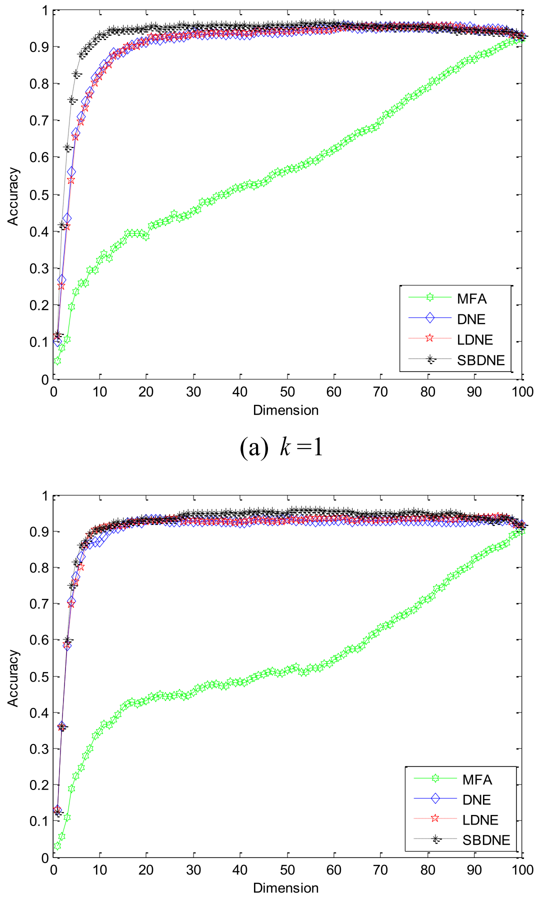
In this experiment, owing to the high dimensionality of ORL dataset, we would reduce dimensionality with two times so as to get a high running speed. Additionally, PCA is employed firstly to reduce the data to be 100 features since it can eliminate the majority of noises. 4 samples of each person in the ORL dataset are selected to be training ones and the rest 6 ones are test ones.

The neighborhood parameter k in MFA, DNE, LDNE and SBDNE is set to be 1 and 3, respectively. We repeat 10 data division for training and test and report the average experimental results in Fig. 5. Although the number of neighbors is different when constructing the adjacent graph, the recognition rates for each method have the consistent tendency. Compared with MFA, DNE and LDNE, SBDNE has a better recognition rate and its optimal discriminative subspace has a relatively low dimensionality so as to reduce the complexity of calculation.

Table II presents the optimal recognition rate and the dimensionality of discriminative sub-space with different number of nearest neighbors. Compared with other methods, SBDNE has not only a better recognition rate but also a lower dimensionality of discriminative sub-space.

When building adjacent graphs, SBDNE not only confirms the locally structure information and the positions of data but also solves the unbalanced problem. By endowing the intra-class data and the inter-class data with different similarities, SBDNE has a more powerful discrimination than MFA, DNE and LDNE so as to make the learned projection matrix be able to more efficiently achieve aggregation among intra-class data and separation among inter-class data.

Fig.4 Samples of face image for the ORL face database



(a) $k=1$

(b) $k=3$

Fig.5 Recognition performance for the ORL database with different neighbor parameter

TabLE II. Performance comparison on the ORL database

	k=1		k =3	
	Sub-dimension	***Recognition rate (%)***	***Sub-dimension***	***Recognition rate (%)***
MFA	100	92.08 ± 0.32	100	89.86 ± 0.33
DNE	62	95.42 ± 1.21	50	93.33 ± 2.19
LDNE	73	95.56 ± 0.17	84	93.89 ± 0.64
SBDNE	53	96.25 ± 0.38	52	95.83 ± 0.43

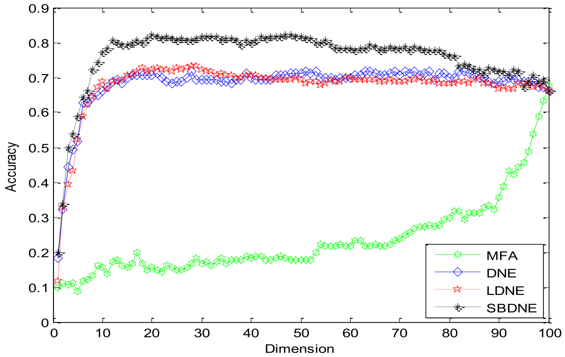
D. Yale Dataset

Yale dataset [17] contains 165 face images of 15 persons, with 11 images for each person. The face expression and light condition for each image are as follows: centered light, with glasses or without glasses, happy, normality, left side light, right side light, sad, sleep, surprise and blink. The size of each image is 32x32 with grayscale from 0 to 255. Fig. 6 shows some face images with different conditions from the Yale dataset.

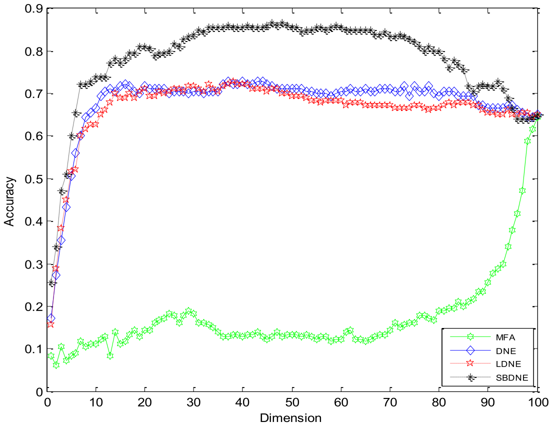
Being similar to ORL face dataset, Yale face samples are also reduced to 100 features via PCA, after which we would respectively employ the methods of MFA, DNE, LDNE and SBDNE to achieve the second dimensionality reduction. Here, we mainly focus on the effect of the number of samples on the recognition rate and all the recognition rates are the average values of 100 experiments with $k=1$. In the experiment, we randomly select 5 (or 7) samples of one person as our training set and the rest ones as test set.

Fig. 7 shows the recognition rates of each discriminative subspace with the different number of samples. From Fig. 7(a) and 7(b), we can see that although the number of samples is different, the general trends of recognition rate are the same. Compared with MFA, DNE and LDNE, SBDNE always presents a better recognition rate and tends to the high recognition rate in a relatively fast speed. Table III provides the optimal recognition rates for the four methods with the different number of samples.

Fig.6 Samples of face image for the Yale face database



(a) 5 training samples



(b) 7 training samples

Fig.7 Recognition performance for the Yale database with different training samples.

TABLE III. PERFORMANCE COMPARISON ON THE YALE DATABASE

	Five training samples		**Seven training samples**	
	Sub-dimension	***Recognition rate (%)***	***Sub-dimension***	***Recognition rate (%)***
MFA	100	67.78 ± 0.34	100	64.78 ± 0.34
DNE	51	71.67 ± 1.12	40	72.18 ± 1.12
LDNE	28	73.33 ± 2.63	38	72.78 ± 2.63
SBDNE	20	82.22 ± 0.64	46	86.67 ± 0.64

E. Yale Dataset

The FERET database is a standard database for evaluating state-of-art face recognition algorithms. In this experiment, a subset, this contains 1400 face images of 200 individuals with 7 images per individual. Fig. 8 shows Sample images for one individual of the FERET subset

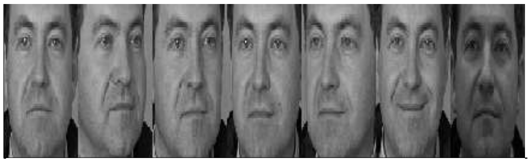


Fig.8 Sample images for one individual of the FERET subset

Being similar to ORL and Yale face dataset, FERET face samples are also reduced to 100 features via PCA, after which we would respectively employ the methods of MFA, DNE, LDNE and SBDNE to achieve the second dimensionality reduction. And also we mainly focus on the effect of the number of samples on the recognition rate and all the recognition rates are the average values of 100 experiments with k =1. In the experiment, we randomly select 3(or 4) samples of one person as our training set and the rest ones as test set.

Table IV provides the optimal recognition rates for the four methods with the different number of samples.

TABLE IV. PERFORMANCE COMPARISON ON THE FERET DATABASE

	Three training samples		**Four training samples**	
	Sub-dimension	***Recognition rate (%)***	***Sub-dimension***	***Recognition rate (%)***
MFA	100	35.75 ± 3.34	100	47.75 ± 2.34
DNE	21	50.12 ± 3.15	38	65.50 ± 1.25
LDNE	25	52.88 ± 2.48	38	66.75 ± 0.95
SBDNE	20	79.37 ± 0.28	22	83.75 ± 0.80

V. CONCLUSION

This paper proposes a new linear dimensionality reduction method, called Similarity-balanced Discriminant Neighborhood Embedding (SBDNE). Based on the MFA and LDNE, SBDNE gives some improvements in that the information of samples' positions and the balanced relationship between the intra-class data and the inter-class data are taken into account. As a result, we are able to give an overall consideration on the preservation of manifold's original geometric structure and utility of classification information.

Through numerical experiments, the advantages of SBDNE are verified on the synthetic dataset and the two face image datasets. By directly using the constructed low-dimensional model, it is able to quickly get the low-dimensional information of new test example, at the same time, with a rising recognition rate. However, SBDNE is still a linear method, so to improve on the classification performance, in the future, we would try to extend it to be non-linear.

APPENDIX A

First, SBDNE computes the similarity function $G\left(\mathbf{x}_i, \mathbf{x}_j\right)$ according to the labels, from which we can know that the similarity of the samples with the same class being farthest from one example is the minimal among the samples with the same class, and the similarity of the samples with the different classes being nearest to one example is the maximal among the samples with the different classes.

According to $G\left(\mathbf{x}_i, \mathbf{x}_j\right)$, the intra-class structure graph $\mathbf{F}^w$ and the inter-class structure graph $\mathbf{F}^b$ are constructed by (10) and (12), respectively. The intra-class compactness is given by

$$\Phi(\mathbf{P})=\sum_{i,j} \| \mathbf{P}^T\mathbf{x}_i - \mathbf{P}^T\mathbf{x}_j \|^2 \; F_{ij}^w$$

and the inter-class divergence is given by

$$\Omega(\mathbf{P})=\sum_{i,j} \| \mathbf{P}^T\mathbf{x}_i - \mathbf{P}^T\mathbf{x}_j \|^2 \; F_{ij}^b \quad .$$

For SBDNE method, the difference $\Psi(\mathbf{P})$ between the sum of the distances between the nearest samples with the different classes and the sum of the distances between the farthest samples with the same class is maximal and its derivation process is as follows:

$$\begin{aligned}\Psi(\mathbf{P}) &= \Phi(\mathbf{P}) - \Omega(\mathbf{P}) \\ &= 2tr\{\mathbf{P}^T\mathbf{X}(\mathbf{D}^b - \mathbf{F}^b)\mathbf{X}^T\mathbf{P} - 2\mathbf{P}^T\mathbf{X}(\mathbf{D}^w - \mathbf{F}^w)\mathbf{X}^T\mathbf{P}\} \\ &= 2tr\{\mathbf{P}^T\mathbf{X}(\mathbf{D}^b - \mathbf{F}^b - \mathbf{D}^w + \mathbf{F}^w)\mathbf{X}^T\mathbf{P}\} \\ &= 2tr\{\mathbf{P}^T\mathbf{X}\mathbf{U}\mathbf{X}^T\mathbf{P}\} \\ &= 2\sum_{i=1}^{d}\mathbf{P}_i^T\mathbf{X}\mathbf{U}\mathbf{X}^T\mathbf{P}\end{aligned}$$

where $\mathbf{U} = \mathbf{D}^b - \mathbf{F}^b - \mathbf{D}^w + \mathbf{F}^w$, according to the method for DNE algorithm, then we have the optimization problem as follows:

$$\begin{cases}\max\limits_{\mathbf{P}} \; tr(\mathbf{P}^T\mathbf{X}\mathbf{U}\mathbf{X}^T\mathbf{P}) \\ s.t. \qquad \mathbf{P}^T\mathbf{P} = \mathbf{I}\end{cases}$$

whose corresponding problem is

$$\max \sum_{i=1}^{d}\mathbf{p}^T{}_i\mathbf{X}\mathbf{U}\mathbf{X}^T\mathbf{p}_i = \sum_{i=1}^{d}\lambda_i \quad .$$

Suppose the eigenvalues of $\mathbf{X}\mathbf{U}\mathbf{X}^T$ are $\lambda_1 \geq \cdots \geq \lambda_d$, we select the r eigenvectors corresponding to the first r eigenvalues to form the transformation matrix, or $\mathbf{P} = [\mathbf{p}_1, \cdots, \mathbf{p}_r]$.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 61373093 and 61033013, by the Natural Science Foundation of Jiangsu Province of China under Grant Nos. BK2011284 and BK201222725, by the Natural Science Foundation of the Jiangsu Higher Education Institutions of China under Grant No.13KJA520001, and by the Qing Lan Project.

References

[1] I. Joliffe. Principal Component Analysis. Springer, New York, 1986

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

[3] K. Fukunnaga, Introduction to Statistical Pattern Recognition, Academic Press, second edition, 1991.

[4] J. B. Tenenbaum, V. D. Silva, J. C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction", Science, 2000, vol. 290, pp. 2319-2323

[5] S. Roweis, L. Saul. "Nonlinear dimensionality reduction by locally linear embedding", Science, 2000, vol. 290, pp. 2323-2326

[6] M. Belkin, P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation", Neural Computation, 2003, vol. 15 pp. 1373-1396

[7] X. F. He, S. C. Yan, Y. X. Hu, P. Niyogi, H. J. Zhang, "Face recognition using Laplacian faces", IEEE Trans. Pattern Analysis and Machine Intelligence, 2005, Volume. 27, pp. 328-340.

[8] X. F. He, D. Cai, S. C. Yan and H. J. Zhang, "Neighborhood preserving embedding", In: Proceedings of IEEE International Conference on Computer Vision, 2005, vol. 2, pp.1208-1213

[9] S. C. Yan, D. Xu, B. Y. Zhang, H. J. Zhang, "Graph Embedding: A General Framework for Dimensionality Reduction", Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

[10] S. C. Yan, D. Xu, B. Y. Zhang, H. J. Zhang, S. Lin, Q. Yang, "Graph Embedding and Extensions: A General Framework for Dimensionality Reduction". IEEE Trans. Pattern Analysis and Machine Intelligence29, 2007, pp. 40-51

[11] Y. Q. Lu, C. Lu, M. Qi, S. Y. Wang, "A Supervised Locality Preserving Projections Based Local Matching Algorithm for Face Recognition", Advances in Computer Science and Information Technology, 2010, Volume. 6059, pp. 28-37

[12] Q. B. You, N. N. Zheng, S. Y. Du, Y. Wu, "Neighborhood discriminant projection for face recognition", Pattern Recognition, 2007, Volume. 28, pp. 1156-1163

[13] W. Zhang, X. Y. Xue, H. Lu, Y. F. Guo, "Discriminant neighborhood embedding for classification", Pattern Recognition, 2006, Volume. 39, pp. 2240-2243.

[14] J. P. Gou, Z. Yi, "Locality-Based Discriminant Neighborhood Embedding", The Computer Journal, 2013, Volume. 56, pp. 1063-1082

[15] W. Zhang, X. Y. Xue, Study on Feature Transformation Algorithm based on K-Nearest-neighbor Classification Rule, Fudan University, China,2007, pp.37-40.

[16] The Database of Faces, Available: http://www.cl.cam.ac.uk/research /dtg/attarchive/facedatabase.html,(accessed December 21, 2013)

[17] UCSD Computer Vision, Available: http://vision.ucsd.edu/content/yale-face-database, (accessed December 21, 2013)

[18] X. R. Li, T. Jiang, K. S. Zhang, "Efficient and robust feature extraction by maximum margin criterion", Neural Networks, 2006, Volume. 17, pp. 157-165