Semi-Supervised Discriminant Analysis using Robust Path-Based Similarity

Yu Zhang & Dit-Yan Yeung
Department of Computer Science and Engineering, Hong Kong University of Science and Technology
{zhangyu,dyyeung}@cse.ust.hk

Abstract

Linear Discriminant Analysis (LDA), which works by maximizing the within-class similarity and minimizing the between-class similarity simultaneously, is a popular dimensionality reduction technique in pattern recognition and machine learning. In real-world applications when labeled data are limited, LDA does not work well. Under many situations, however, it is easy to obtain unlabeled data in large quantities. In this paper, we propose a novel dimensionality reduction method, called Semi-Supervised Discriminant Analysis (SSDA), which can utilize both labeled and unlabeled data to perform dimensionality reduction in the semi-supervised setting. Our method uses a robust path-based similarity measure to capture the manifold structure of the data and then uses the obtained similarity to maximize the separability between different classes. A kernel extension of the proposed method for nonlinear dimensionality reduction in the semi-supervised setting is also presented. Experiments on face recognition demonstrate the effectiveness of the proposed method.

1. Introduction

Linear Discriminant Analysis (LDA) [13, 21] is a popular dimensionality reduction technique in pattern recognition and machine learning. It aims to maximize the within-class similarity while minimizing the between-class similarity simultaneously. LDA is widely used in many applications, such as face recognition, speech recognition, and character recognition, due to its effectiveness and computational efficiency. However, in some applications, LDA encounters the so-called *small sample size* (SSS) problem [11] which arises when the sample size is much smaller than the dimensionality of the feature space. The performance of LDA will seriously deteriorate under such situations because there may not be enough data to make the within-class scatter matrix nonsingular. Several methods have been proposed to overcome the SSS problem, e.g., PseudoLDA [18], PCA+LDA [1], Direct-LDA [27], NullLDA [11], and DualLDA [24]. PseudoLDA overcomes the singularity problem by substituting the inverse of the within-class scatter matrix with its pseudo-inverse. PCA+LDA first applies PCA [17] to project the data into a lower-dimensional space so that the within-class scatter matrix computed there is nonsingular, and then applies LDA in the lower-dimensional space. Direct-LDA projects data into the range space of the between-class scatter matrix by diagonalizing the between-class scatter matrix and then minimizes the within-class scatter in the reduced space. NullLDA first projects the data to the null space of the within-class scatter matrix and then maximizes the between-class scatter in this space. It is similar to the Discriminative Common Vectors method [6]. DualLDA, which combines the ideas from PCA+LDA and NullLDA, applies LDA in the range space and the null space of the within-class scatter matrix separately and then integrates the two parts together to get the final transformation. There also exists another approach to address the SSS problem, with 2DLDA [26] being the representative of this approach. The major difference between 2DLDA and the above algorithms is in the data representation. Specifically, 2DLDA works on data represented as matrices instead of vectors so that the dimensionality of the data representation can be kept small to avoid the SSS problem. Moreover, in some applications such as face and object recognition, 2DLDA can preserve the spatial information in an image which may be useful for classification.

In many real-world applications, labeled data are hard or expensive to obtain because laborious human labeling effort is required. On the other hand, abundant supply of unlabeled data is available at very low cost. In recent years, semi-supervised learning has emerged as a hot topic within the machine learning research community. One common form of semi-supervised learning is to utilize unlabeled data to aid classification or regression tasks when labeled data are scarce. A good survey of semi-supervised learning methods can be found in [29]. Some early semi-supervised learnng methods include Co-Training [4] and Transductive SVM [3]. More recently, graph-based semi-supervised learning methods [2, 28, 30] have aroused the interests of many researchers. These methods model the relationships between data points in the form of a graph,

in which label information from the labeled data points is propagated to the unlabeled data points through the graph.

This leads us to ask the following question: *Can unlabeled data be utilized to help LDA to alleviate the SSS problem?* In this paper, we propose a novel semi-supervised dimensionality reduction algorithm called Semi-Supervised Discriminant Analysis (SSDA). Even though there already exists another semi-supervised LDA algorithm called SDA [5] which exploits the local neighborhood information of data points in performing dimensionality reduction, our SSDA algorithm exploits the global structure of the data and is robust against noise in defining the neighborhood relationships. SSDA first constructs a graph using a robust path-based similarity measure to capture the manifold structure of the data. Unlike some existing graph-based semi-supervised learning methods which make use of the similarity or affinity matrix to define a manifold-based regularization term for an optimization problem formulated under the regularization framework, we propose a new optimality criterion for LDA by exploiting the interplay between labeled and unlabeled data. Like the original LDA algorithm, learning in SSDA also reduces to solving a generalized eigenvalue problem to obtain the projection directions for dimensionality reduction.

The rest of this paper is organized as follows. We first briefly review the traditional LDA algorithm in Section 2. We then present our SSDA algorithm in Section 3. Section 4 reports some experimental results based on two commonly used face databases to demonstrate the effectiveness of our method. Finally, we conclude our paper in the last section.

2. Background

Given a training set of n data points, $D = \{x_1, \ldots, x_n\}$ where $x_i \in \mathbb{R}^N$ $(i = 1, \ldots, n)$, LDA tries to find a projection matrix W^* that maximizes the ratio of the trace of the between-class scatter matrix S_b to that of the within-class scatter matrix S_w:

$$W^* = \arg\max_W \frac{\mathrm{trace}(W^T S_b W)}{\mathrm{trace}(W^T S_w W)}. \tag{1}$$

Let the data set D be partitioned into $C \geq 2$ disjoint classes Π_i $(i = 1, \ldots, C)$ where class Π_i contains n_i examples. The scatter matrices S_b and S_w are defined as:

$$S_b = \sum_{k=1}^{C} n_k (\bar{m}_k - \bar{m})(\bar{m}_k - \bar{m})^T \tag{2}$$

$$S_w = \sum_{k=1}^{C} \sum_{x_i \in \Pi_k} (x_i - \bar{m}_k)(x_i - \bar{m}_k)^T \tag{3}$$

where $\bar{m} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean of the whole data set D and $\bar{m}_k = \frac{1}{n_k}\sum_{x_i \in \Pi_k} x_i$ is the class mean of Π_k.

W^* can be computed from the eigenvectors of $S_w^{-1} S_b$, where S_w^{-1} denotes the matrix inverse of S_w [15]. According to [25], W^* computed as above may not be optimal with respect to the optimality criterion in Eq. (1), but it is a computationally simple and good approximate solution sufficient for many applications. Thus many applications still use this approach to obtain the solution.

Fukunnaga [15] proved that W^* can also be computed by the simultaneous diagonalization of S_w and S_b. Finally, W^* satisfies ${W^*}^T S_w W^* = I_t$ and ${W^*}^T S_b W^* = \mathrm{diag}\{\lambda_1(S_w^{-1} S_b), \ldots, \lambda_t(S_w^{-1} S_b)\}$, where $\lambda_i(B)$ is the ith largest eigenvalue of matrix B, I_t is the identity matrix of size $t \times t$, and $\mathrm{diag}\{d_1, \ldots, d_t\}$ is a $t \times t$ diagonal matrix whose (i, i) element is d_i.

3. Our Semi-Supervised Discriminant Analysis Algorithm

3.1. Robust Path-Based Similarity Measure

We denote a set of n points in some multidimensional Euclidean space by $D = \{x_1, \ldots, x_n\}$. This data set can be represented by an undirected graph $G = (V, E)$, with the vertex set $V = \{1, \ldots, n\}$ corresponding to the data points in D and the edge set $E \subseteq V \times V$ representing the relationships between data points. Each edge is assigned a weight w_{ij} which reflects the similarity between points x_i and x_j:

$$w_{ij} = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) & \text{for } i \neq j \\ 0 & \text{for } i = j. \end{cases} \tag{4}$$

The scaling parameter σ controls how fast w_{ij} decreases with the distance between x_i and x_j.

The pairwise similarity w_{ij} defined above is only determined by the Euclidean distance between x_i and x_j. It cannot reveal whether the two points belong to the same class. To capture this information, we exploit the underlying manifold structure of the whole data set based on a robust path-based similarity measure as described below.

We first define a path-based similarity measure as in our previous work [7, 8]. Let $\mathcal{P}_{ij}$ denote the set of all paths connecting vertices i and j. For each path $p \in \mathcal{P}_{ij}$, the effective similarity s_{ij}^p is the minimum edge weight along the path. The path-based similarity measure s'_{ij} is defined as the maximum effective similarity among all paths in $\mathcal{P}_{ij}$:

$$s'_{ij} = \max_{p \in \mathcal{P}_{ij}} \left\{ \min_{1 \leq h < |p|} w_{p[h]p[h+1]} \right\}, \tag{5}$$

where $p[h]$ denotes the hth vertex along path p and $|p|$ denotes the number of vertices in p.

According to [7, 8], this similarity is sensitive to noise and outliers. We proposed a robust estimation approach to

compute the weight α_i' for each point x_i as:

$$\alpha_i' = \sum_{x_j \in \mathcal{N}_i} \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right),$$

where $\mathcal{N}_i$ denotes the neighborhood of x_i. To make the weights insensitive to σ, normalized weights are computed as $\alpha_i = \alpha_i' / \max_{x_i \in D} \alpha_i'$.

Finally, the robust path-based similarity measure is expressed as:

$$s_{ij} = \max_{p \in \mathcal{P}_{ij}} \left\{ \min_{1 \le h < |p|} \alpha_{p[h]} \alpha_{p[h+1]} w_{p[h]p[h+1]} \right\}. \quad (6)$$

3.2. Objective Function for SSDA

Suppose we have l labeled examples $x_1, x_2, \ldots, x_l \in \mathbb{R}^N$ with class labels from C classes $\Pi_i (i = 1, \ldots, C)$ and m unlabeled examples $x_{l+1}, x_{l+2}, \ldots, x_{l+m} \in \mathbb{R}^N$ with unknown class memberships. So there are a total of $n = l + m$ examples and usually $l \ll m$. When l is too small compared with the input dimensionality, LDA generally does not perform very well. To remedy this problem, we want to incorporate unlabeled data to improve the performance of LDA. We propose to maximize a new objective function as follows:

$$\max_w \frac{w^T S_b w}{w^T S_w w + J(w)}, \quad (7)$$

where w is an N-dimensional projection vector, S_b and S_w are the between-class and within-class scatter matrices as defined in Eqs. (2) and (3), and $J(w)$ is a regularization term which is learned from both the labeled and unlabeled data.

We first construct the robust path-based similarity matrix S using Eq. (6) based on the whole data set. In the following, we will describe how to compute $J(w)$ using S.

First, we divide S into four blocks:

$$S = \left(\begin{array}{c|c} S^{ll} & S^{lu} \\ \hline S^{ul} & S^{uu} \end{array}\right),$$

where S^{ll} captures the similarity between labeled data points, S^{lu} and S^{ul} capture the similarity between labeled and unlabeled data points with $S^{lu} = (S^{ul})^T$, and S^{uu} captures the similarity between unlabeled data points. Because ground-truth label information already exists for the labeled data and the similarity information S^{ll} may not be fully in line with the label information, we choose to discard S^{ll}.

Since S^{lu} and S^{ul} contain the same information, we just need to use one of them, say S^{ul}. Recall that the optimality criterion of LDA favors having data points from the same class to be close to their class mean. Following this idea, if an unlabeled data point is similar to some labeled data point from the ith class, then we expect the unlabeled data point to be close to the class mean of the ith class. We define the similarity Q_{ij} between the ith unlabeled data point x_{l+i} and the jth class as $Q_{ij} = \max_{x_t \in \Pi_j} \{S_{it}^{ul}\}$, where S_{it}^{ul} is (i,t) element of S^{ul}. That is, if an unlabeled data point x_{l+i} has higher similarity to some labeled point from the jth class, then x_{l+i} is more likely to belong to the jth class. Let Q denote the similarity matrix between unlabeled data points and class means with Q_{ij} being its elements. Similar to [16], a Laplacian-style measure is defined as follows:

$$\begin{aligned} & J_1(w) \\ = & \sum_{i=1}^{m}\sum_{j=1}^{C}(w^T x_{l+i} - w^T \bar{m}_j)^2 Q_{ij} \\ = & w^T \left[\sum_{i=1}^{m}\left(\sum_{j=1}^{C} Q_{ij}\right) x_{l+i} x_{l+i}^T + \sum_{j=1}^{C}\left(\sum_{i=1}^{m} Q_{ij}\right) \bar{m}_j \bar{m}_j^T \right. \\ & \left. -2\sum_{i=1}^{m}\sum_{j=1}^{C} Q_{ij} x_{l+i} \bar{m}_j^T \right] w \\ = & w^T (X_u D_1 X_u^T + M D_2 M^T - 2 X_u Q M^T) w \\ = & w^T (X_u D_1 X_u^T + M D_2 M^T - X_u Q M^T - M Q^T X_u^T) w \\ = & w^T L_1 w, \end{aligned} \quad (8)$$

where $X_u = [x_{l+1}, \ldots, x_{l+m}]$, $\bar{m}_i$ is the class mean of the ith class, $M = [\bar{m}_1, \ldots, \bar{m}_C]$, and D_1 and D_2 are diagonal matrices whose entries are the row sums and column sums of Q, respectively. The second last step in Eq. (8) is just to make L_1 symmetrical to facilitate subsequent processing.

Next, we discuss how to utilize S^{uu}. If two points have high similarity, we expect them to be close to each other in the reduced space. Thus, a Laplacian-style measure can be defined as follows:

$$\begin{aligned} & J_2(w) \\ = & \sum_{ij}(w^T x_{l+i} - w^T x_{l+j})^2 S_{ij}^{uu} \\ = & 2w^T \left[\sum_i \left(\sum_j S_{ij}^{uu}\right) x_{l+i} x_{l+i}^T - \sum_{ij} S_{ij}^{uu} x_i x_j^T\right] w \\ = & 2w^T X_u (D^{uu} - S^{uu}) X_u^T w \\ = & 2w^T X_u L_2 X_u^T w, \end{aligned} \quad (9)$$

where S_{ij}^{uu} is (i,j) element of S^{uu}, D^{uu} is a diagonal matrix whose entries are the column sums of S^{uu} and $L_2 = D^{uu} - S^{uu}$ is the Laplacian matrix [12] of S^{uu}.

Finally, we combine Eqs. (8) and (9) to get the objective function for the optimization problem of our SSDA algorithm and maximize it with respect to w:

$$\max_w \frac{w^T S_b w}{w^T (S_w + L_1 + \alpha X_u L_2 X_u^T) w}, \quad (10)$$

where α is a control parameter.

According to [15], solving this optimization problem is equivalent to solving the following generalized eigenvalue

problem:

$$S_b w = \lambda(S_w + L_1 + \alpha X_u L_2 X_u^T)w. \quad (11)$$

When the number of data points is smaller than the dimensionality of the data, $S_w + L_1 + \alpha X_u L_2 X_u^T$ in Eq. (11) may be singular and hence the eigen-decomposition problem becomes unstable. To avoid this problem, we adopt the idea of Tikhonov regularization as in regularized discriminant analysis [14]. So the generalized eigenvalue problem in Eq. (11) becomes:

$$S_b w = \lambda(S_w + L_1 + \alpha X_u L_2 X_u^T + \beta I)w, \quad (12)$$

where $\beta > 0$ and I is the identity matrix.

3.3. The Algorithm

The SSDA algorithm can be summarized as follows:

1. Construct the robust similarity matrix S using Eq. (6).
2. Construct the scatter matrices S_b and S_w defined in Eqs. (2) and (3) using only labeled data.
3. Construct the graph Laplacian matrices L_1 and L_2 for the regularization terms using Eqs. (8) and (9).
4. Solve the generalized eigenvalue problem in Eq. (12). Since the rank of S_b is at most $C-1$, we have $C-1$ eigenvectors, denoted as $w_1, \ldots, w_{C-1}$, corresponding to the nonzero eigenvalues.
5. Let $W = [w_1, w_2, \ldots, w_{C-1}]$. Data points can be embedded into the lower-dimensional space via the following transformation: $x \to y = W^T x$.

3.4. Discussions

As pointed out in [10], the similarity measure defined in Eq. (5) is a density-sensitive similarity measure. So our method can be regarded as adopting the so-called *cluster assumption* in semi-supervised learning [9, 10], which says that two points are likely to have the same class label if there exists a path connecting them by passing through regions of high density only. Here we interpret the cluster assumption in a somewhat different way. Specifically, if two points are connected by a path in which adjacent vertices have high similarity, then the two points will likely belong to the same class.

SDA [5] is also a semi-supervised discriminant analysis method which makes use of both labeled and unlabeled data. Unlike our method, however, SDA adopts the so-called *manifold assumption* [28], which says that nearby points have similar low-dimensional representations and use neighborhood information to find better embedding. However, in situations where the intra-class variance is larger than the inter-class variance, the neighborhood information may be noisy and the performance of SDA will deteriorate. Our experimental results confirm this speculation. In contrast to SDA, our method exploits the global manifold structure which will not be affected by noisy neighborhood. In general, the manifolds can be nonlinear and elongated in structure. With the help of the path-based similarity measure, we can convert elongated manifolds into compact ones which can help further classification.

3.5. Kernel SSDA

The SSDA algorithm presented above is a linear method. In general, it may fail to handle data with nonlinear manifold structure. In this subsection, we discuss how to extend SSDA to a reproducing kernel Hilbert space (RKHS) which corresponds to a feature space.

We consider the RKHS or feature space $\mathcal{F}$ induced by a nonlinear mapping $\phi : \mathbb{R}^n \to \mathcal{F}$. For a proper ϕ, the inner product operation $\langle \cdot, \cdot \rangle$ in $\mathcal{F}$ can be defined as some positive semi-definite kernel function $K(\cdot, \cdot)$ such that $\langle \phi(x), \phi(y) \rangle = K(x, y)$. Some popular kernel functions are: Gaussian kernel $K(x, y) = \exp(-\|x-y\|^2/\sigma^2)$; polynomial kernel $K(x, y) = (1 + x^T y)^d$.

Suppose $\phi(\bar{m}) = 0$, where $\bar{m}$ is the sample mean of the labeled data; otherwise, we can apply a centering transform to make $\phi(\bar{m}) = 0$. Let $\Phi_l = [\phi(x_1), \ldots, \phi(x_l)]$ be the labeled data matrix in $\mathcal{F}$, $\Phi_u = [\phi(x_{l+1}), \ldots, \phi(x_n)]$ be the unlabeled data matrix in $\mathcal{F}$, $\Phi = [\Phi_l, \Phi_u] = [\phi(x_1), \ldots, \phi(x_n)]$ be the total data matrix in $\mathcal{F}$, and $\Psi = [\phi(\bar{m}_1), \ldots, \phi(\bar{m}_C)]$ be the class mean matrix in $\mathcal{F}$. The between-class scatter matrix S_b^ϕ in $\mathcal{F}$ can be calculated as $S_b^\phi = \Psi D \Psi^T$, where D is a diagonal matrix whose (i, i) element is n_i. The within-class scatter matrix S_w^ϕ in $\mathcal{F}$ is defined as $S_w^\phi = S_t^\phi - S_b^\phi = \Phi_l \Phi_l^T - \Psi D \Psi^T$, where S_t^ϕ is the total scatter matrix on labeled data in $\mathcal{F}$.

The first regularization term J_1^ϕ can be formulated as

$$J_1^\phi = \Phi_u D_1 \Phi_u^T + \Psi D_2 \Psi^T - \Phi_u Q \Psi^T - \Psi Q^T \Phi_u^T.$$

So the problem in Eq. (11) in $\mathcal{F}$ can be written as follows:

$$\Psi D \Psi^T v = \lambda(\Phi_l \Phi_l^T - \Psi D \Psi^T + J_1^\phi + \alpha \Phi_u L_2 \Phi_u^T)v. \quad (13)$$

From the analysis of [20], the eigenvector of Eq. (13) is a linear combination of $\phi(x_1), \ldots, \phi(x_n)$. Thus there exist coefficients $\gamma_i, i = 1, \ldots, n$ such that $v = \sum_{i=1}^n \gamma_i \Phi(x_i) = \Phi a$, where $a = (\gamma_1, \ldots, \gamma_n)^T$.

With some algebraic calculations, we can get

$$\begin{aligned} & \Psi D \Psi^T v = \lambda(\Phi_l \Phi_l^T - \Psi D \Psi^T + J_1^\phi + \alpha \Phi_u L_2 \Phi_u^T)v \\ \Rightarrow \quad & \Psi D \Psi^T \Phi a = \lambda(\Phi_l \Phi_l^T - \Psi D \Psi^T + J_1^\phi + \alpha \Phi_u L_2 \Phi_u^T)\Phi a \\ \Rightarrow \quad & \Phi^T \Psi D \Psi^T \Phi a = \lambda \Phi^T (\Phi_l \Phi_l^T - \Psi D \Psi^T + J_1^\phi \\ & + \alpha \Phi_u L_2 \Phi_u^T)\Phi a. \end{aligned}$$

Figure 1. Sample images for one subject in the CMU PIE face database. For each subject, there are about 49 frontal face images taken under different illumination conditions.

Since $\Phi^T\Psi$, $\Phi^T\Phi_l$, $\Phi^T\Phi_u$, and $\Phi^T J_1^\phi \Phi$ can be calculated by applying the kernel function K, the generalized eigenvalue problem can be solved without knowing the explicit form of the mapping ϕ. Let the column vectors $a_1, \ldots, a_{C-1}$ be the eigenvectors corresponding to the nonzero eigenvalues and $\Upsilon = [a_1, \ldots, a_{C-1}]$ be the transformation matrix. Then a data point x can be projected into the lower-dimensional space via the following transformation: $x \to y = \Upsilon^T\Phi^T\phi(x) = \Upsilon^T\mathcal{K}$, where $\mathcal{K} = [K(x_1, x), \ldots, K(x_n, x)]^T$.

4. Experiments

In this section, we report some experimental results based on two face databases to evaluate the performance of SSDA.

4.1. Experimental Setup

Subspace-based methods have achieved great successes in many face recognition applications [23, 1]. Previous research found that face images usually lie in some low-dimensional subspace within the ambient image space. Two famous methods are Eigenface [23] (based on PCA) and Fisherface [1] (based on LDA). Many variants have also been proposed. These subspace methods use different dimensionality reduction techniques to obtain a low-dimensional subspace and then perform classification in the subspace using some classifier. Unlike previous methods, SDA and our method, SSDA, are semi-supervised subspace methods derived from LDA but use both labeled and unlabeled data for training. In our experiments, we compare SSDA with several subspace methods, including Eigenface, Fisherface, and SDA. After dimensionality reduction has been performed, we use a simple nearest-neighbor classifier to perform classification in the subspace. Moreover, we also compare SSDA with the baseline method which simply uses the nearest-neighbor classifier in the original image space. For Fisherface, we use PCA to preserve 95% variance of the data. The parameter β in Eq. (12) is fixed to 10^{-3} for SSDA and so is SDA.

4.2. PIE Face Database

We use the PIE face database [22] for the first set of experiments. The database contains 41,368 face images from 68 individuals. The face images were captured under varying pose, illumination and expression conditions. For our experiments, we choose the frontal pose (C27) with varying lighting and illumination conditions. There are about 49 images for each subject. Before the experiments, we resize each image to a resolution of 32×32 pixels. Some sample images are shown in Figure 1.

In the first experiment, 30 images are randomly selected for each person to form the training set and the rest to form the test set. Of the 30 images for each person, one image is randomly selected and labeled while the other 29 images remain unlabeled. We perform 20 random splits and report the average results over the 20 trials. Table 1 reports the recognition rates of different methods evaluated on the unlabeled training data and the test data separately. Because there is only one labeled training example per person, Fisherface cannot work because the within-class scatter matrix is a zero matrix. The baseline method does not consider the manifold structure of data and Eigenface, an unsupervised method, does not utilize the labeled data, and hence both methods get poor results. On the other hand, both SDA and SSDA exploit the manifold structure and the label information and hence get better results. Moreover, SSDA achieves the best results among all methods tested.

Table 1. Recognition error rates (in mean±std-dev) on PIE when there are one labeled and 29 unlabeled examples.

Method	Unlabeled error	Test error
Baseline	0.7523±0.0146	0.7579±0.0150
Eigenface	0.7874±0.0131	0.7935±0.0148
Fisherface	-	-
SDA	0.6016±0.0372	0.6032±0.0330
SSDA	**0.5341±0.0319**	**0.5403±0.0333**

In the second experiment, the settings are almost the same as the first one. The only difference is that two images are randomly selected and labeled leaving the other 28 images unlabeled. Table 2 reports the results. It can be seen that SSDA gives better result than Fisherface, implying that unlabeled data can help LDA when there are only very few labeled examples. SSDA again achieves the best result among all methods. To our surprise, SDA is worse than Fisherface. This is because SDA treats the points in the neighborhood as equally important, which can bring some

Figure 2. Sample images for one subject in the AR face database. The images in the first and second rows were taken in different sessions.

Figure 3. Recognition error rates of SSDA on the unlabeled training data and the test data of PIE under varying α values.

noise to deteriorate the performance. Moreover, we also investigate the effect of parameter α in Eq. (12) on the performance of SSDA. The recognition error rates on the unlabeled training data and the test data are plotted in Figure 3. We can see that when α varies in the range $[0.1, 1.1]$, the performance of SSDA only changes slightly for both data sets, with the maximum percentage change being 0.46% for the unlabeled training data and 0.77% for the test data. This shows the relative insensitivity of α and hence it is easy to choose an appropriate value for SSDA to deliver good performance.

Table 2. Recognition error rates (in mean±std-dev) on PIE when there are two labeled and 28 unlabeled examples.

Method	Unlabeled error	Test error
Baseline	0.6155±0.0171	0.6278±0.0150
Eigenface	0.6556±0.0163	0.6658±0.0148
Fisherface	0.3159±0.0254	0.3303±0.0264
SDA	0.4960±0.0282	0.5088±0.0330
SSDA	**0.1809±0.0225**	**0.1972±0.0217**

4.3. AR Face Database

We next use the AR face database [19] for the second set of experiments. The database contains over 4,000 color face images from 126 persons, which include 70 men and 56 women. The face images are all frontal view images with different expressions, illuminations and occlusions. There are 26 images for each person taken in two sessions, each having 13 images. In our experiments, 2,600 images of 100 persons (50 men and 50 women) are used. Before the experiments, each image is converted to gray scale and normalized to 33×24 pixels. Some typical images are shown in Figure 2.

We conduct four experiments on the AR database. For each subject, we randomly select 13 images for the training set and the rest for the test set. Among the 13 images chosen for the training set, we randomly choose $p \in \{2, 3, 4, 5\}$ images and label them. The four experiments correspond to different values of p. For each configuration, we perform 20 random trials and report the average recognition results in Table 3. We can see that SSDA achieves the best results among all methods in all four experiments. To our surprise, SDA is only slightly better than the baseline method and Eigenface but is significantly worse than Fisherface. A possible explanation is that the face images in the AR database have large intra-person (or intra-class) variability due to expression differences and occlusion, so that intra-person variance may be larger than inter-person variance. As a result, nearby points may belong to different classes. Moreover, the face images were taken during two different sessions at different times so that the appearance of the same person may look different, resulting in different distributions of the data points in the two sessions. The neighbors of a data point are more likely to belong to different classes. If data points within the same neighborhood are treated to be from the same class as SDA does, the recognition accuracy may be seriously affected. As in the first set of experiments for the PIE database, we also investigate the effect of parameter α in Eq. (12) on the performance of SSDA. The recognition error rates on the unlabeled training data and the test data are plotted in Figures 4–7. We can see that when α varies in the range $[0.1, 1.1]$, the performance of SSDA does not change very much and the maximum percentage changes for the unlabeled training data and the test data are 2.87% and 2.68% when $p = 2$, 2.21% and 1.91% when $p = 3$, 1.51% and 1.15% when $p = 4$, and 0.93% and 0.72% when $p = 5$. So the performance of SSDA is not very sensitive to α.

Table 3. Recognition error rates (in mean±std-dev) on AR for different p values. TOP LEFT: $p = 2$; TOP RIGHT: $p = 3$; BOTTOM LEFT: $p = 4$; BOTTOM RIGHT: $p = 5$.

Method	Unlabeled error	Test error
Baseline	0.8565±0.0119	0.8517±0.0120
Eigenface	0.8645±0.0116	0.8584±0.0113
Fisherface	0.5259±0.0192	0.5188±0.0173
SDA	0.8158±0.0094	0.7644±0.0144
SSDA	**0.4268±0.0385**	**0.4184±0.0363**

Method	Unlabeled error	Test error
Baseline	0.8110±0.0135	0.8080±0.0144
Eigenface	0.8202±0.0125	0.8155±0.0122
Fisherface	0.4771±0.0265	0.4704±0.0229
SDA	0.7778±0.0136	0.6801±0.0169
SSDA	**0.2820±0.0236**	**0.2849±0.0217**

Method	Unlabeled error	Test error
Baseline	0.7684±0.0106	0.7695±0.0129
Eigenface	0.7768±0.0105	0.7769±0.0118
Fisherface	0.3021±0.0134	0.3014±0.0152
SDA	0.7386±0.0161	0.6105±0.0166
SSDA	**0.2287±0.0113**	**0.2246±0.0149**

Method	Unlabeled error	Test error
Baseline	0.7338±0.0130	0.7354±0.0136
Eigenface	0.7424±0.0129	0.7430±0.0127
Fisherface	0.3209±0.0188	0.3250±0.0150
SDA	0.6941±0.0177	0.5179±0.0210
SSDA	**0.1929±0.0142**	**0.1941±0.0158**

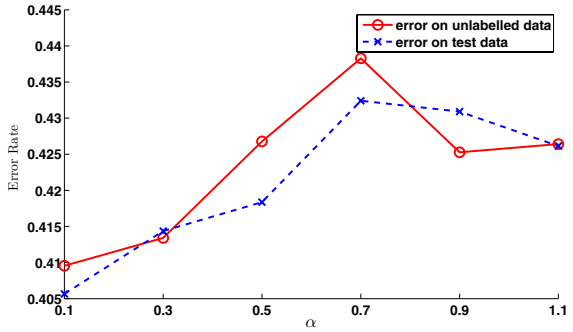

Figure 4. Recognition error rates of SSDA on the unlabeled training data and the test data of AR with $p = 2$ under varying α values.

Figure 5. Recognition error rates of SSDA on the unlabeled training data and the test data of AR with $p = 3$ under varying α values.

Figure 6. Recognition error rates of SSDA on the unlabeled training data and the test data of AR with $p = 4$ under varying α values.

Figure 7. Recognition error rates of SSDA on the unlabeled training data and the test data of AR with $p = 5$ under varying α values.

5. Conclusion

In this paper, we have proposed a new dimensionality reduction algorithm called Semi-Supervised Discriminant Analysis. It can make use of both labeled and unlabeled data in learning a transformation to achieve dimensionality reduction. The similarity between data points is represented by a robust path-based similarity measure so that the global manifold structure of the data can be captured well with high robustness. The global manifold structure plays a crucial role in maximizing the discrimination ability of LDA when labeled training data are very limited. Experiments performed on two commonly used face databases show very promising results when compared with other related methods. In our future research, we will generalize our method to other dimensionality reduction techniques.

Acknowledgements

This research has been supported by Competitive Earmarked Research Grant (CERG) 621407 from the Research Grants Council (RGC) of the Hong Kong Special Administrative Region, China.

References

[1] 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*, 19(7):711–720, 1997.

[2] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

[3] K. Bennett and A. Demiriz. Semi-supervised support vector machines. In *Advances in Neural Information Processing Systems 11*, pages 368–374, Vancouver, British Columbia, Canada, 1998.

[4] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Workshop on Computational Learning Theory*, pages 92–100, Madison, Wisconsin, USA, 1998.

[5] D. Cai, X. He, and J. Han. Semi-supervised discriminant analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, Rio de Janeiro, Brazil, 2007.

[6] H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana. Discriminative common vectors for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(1):4–13, 2005.

[7] H. Chang and D. Y. Yeung. Robust path-based spectral clustering with application to image segmentation. In *Proceedings of the Tenth IEEE International Conference on Computer Vision*, pages 278–285, Beijing, China, 2005.

[8] H. Chang and D. Y. Yeung. Graph Laplacian kernels for object classification from a single example. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2011–2016, New York, NY, USA, 2006.

[9] O. Chapelle, J. Weston, and B. Schölkopf. Cluster kernels for semi-supervised learning. In *Advances in Neural Information Processing Systems 15*, pages 585–592, Vancouver, British Columbia, Canada, 2002.

[10] O. Chapelle and A. Zien. Semi-supervised classification by low density separation. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 57–64, Barbados, 2005.

[11] L. Chen, H. Liao, M. Ko, J. Lin, and G. Yu. A new LDA-based face recognition system which can solve the small sample size problem. *Pattern Recognition*, 33(10):1713–1726, 2000.

[12] F. R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, Rhode Island, 1997.

[13] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[14] J. H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175, 1989.

[15] K. Fukunnaga. *Introduction to Statistical Pattern Recognition*. Academic Press, New York, 1991.

[16] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang. Face recognition using Laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):328–340, 2005.

[17] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1986.

[18] W. J. Krzanowski, P. Jonathan, W. V. McCarthy, and M. R. Thomas. Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data. *Applied Statistics*, 44(1):101–115, 1995.

[19] A. M. Martínez and R. Benavente. The AR-face database. Technical Report 24, CVC, 1998.

[20] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, A. Smola, and K.-R. Müller. Constructing descriptive and discriminative nonlinear features: Rayleigh coefficients in kernel feature spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):623–628, 2003.

[21] C. R. Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statictical Society*, 10:159–203, 1948.

[22] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1615–1618, 2003.

[23] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[24] X. Wang and X. Tang. Dual-space linear discriminant analysis for face recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 564–569, Washington, DC, 2004.

[25] S. Yan and X. Tang. Trace quotient problems revisited. In *Proceedings of the 9th European Conference on Computer Vision*, pages 232–244, Graz, Austria, 2006.

[26] J.-P. Ye, R. Janardan, and Q. Li. Two-dimensional linear discriminant analysis. In *Advances in Neural Information Processing Systems 17*, pages 1529–1536, MIT Press, Cambridge, MA, 2004.

[27] H. Yu and J. Yang. A direct LDA algorithm for high-dimensional data - with application to face recognition. *Pattern Recognition*, 34(10):2067–2070, 2001.

[28] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16*, Vancouver, British Columbia, Canada, 2003.

[29] X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI, 2006.

[30] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the Twentieth International Conference on Machine Learning*, pages 912–919, Washington, DC, 2003.