Neural Comput & Applic (2014) 24:1697–1706
DOI 10.1007/s00521-013-1403-1

ORIGINAL ARTICLE

Graph embedding discriminant analysis for face recognition

Cairong Zhao · Zhihui Lai · Duoqian Miao · Zhihua Wei · Caihui Liu

Received: 24 January 2013 / Accepted: 29 March 2013 / Published online: 13 April 2013

Abstract This paper develops a supervised discriminant technique, called graph embedding discriminant analysis (GEDA), for dimensionality reduction of high-dimensional data in small sample size problems. GEDA can be seen as a linear approximation of a multimanifold-based learning framework in which nonlocal property is taken into account besides the marginal property and local property. GEDA seeks to find a set of perfect projections that not only can impact the samples of intraclass and maximize the margin of interclass, but also can maximize the nonlocal scatter at the same time. This characteristic makes GEDA more intuitive and more powerful than linear discriminant analysis (LDA) and marginal fisher analysis (MFA). The proposed method is applied to face recognition and is examined on the Yale, ORL and AR face image databases. The experimental results show that GEDA consistently outperforms LDA and MFA when the training sample size per class is small.

Keywords Graph embedding · Face recognition · Marginal fisher analysis (MFA) · Manifold learning · Linear discriminant analysis (LDA) · Pattern recognition · Principal component analysis (PCA)

Cairong Zhao and Zhihui Lai contributed equally to this work.

C. Zhao (✉) · D. Miao · Z. Wei · C. Liu
Department of Computer Science and Technology, Tongji University, Shanghai 201804, China
e-mail: cairong.zhao@yahoo.com

C. Zhao · D. Miao · Z. Wei · C. Liu
The Key Laboratory of "Embedded System and Service Computing", Ministry of Education, Shanghai 201804, China

Z. Lai (✉)
Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Harbin 518055, China
e-mail: lai_zhi_hui@163.com

1 Introduction

Dimensionality reduction is to reconstruct a meaningful low-dimensional representation of high-dimensional data. Since there are large volumes of high-dimensional data in numerous real-world applications, dimensionality reduction is a fundamental problem in many scientific fields. From the perspective of pattern recognition, dimensionality reduction is an effective method of avoiding the "curse of dimensionality" [1] and improving the computational efficiency of pattern matching. Therefore, techniques for dimensionality reduction in supervised or unsupervised learning tasks have attracted much attention in computer vision and pattern recognition. Among them, the linear algorithms principal component analysis (PCA) [2, 4] and linear discriminant analysis (LDA) [3, 4] have been the two most popular algorithms because of their relative simplicity and effectiveness. In the past few years, many manifold-based algorithms [5–10, 13–18] have been proposed to discover intrinsic low-dimension embedding of high-dimensional data. A linear technique, locality preserving projections (LPP) [5], has been proposed for dimensionality reduction that can preserve local relationships within the data set that lies on a lower dimensional manifold. Other nonlinear methods, such as isometric feature mapping (ISOMAP) [6], local linear embedding (LLE) [7] and Laplacian Eigenmap [8], have been proposed to find the intrinsic low-dimensional nonlinear data structures hidden in observation space. Two-dimensional local graph embedding discriminant analysis (2DLGEDA) [13] could directly extract the optimal projective vectors directly from

images based on the scatter difference criterion. Wan et al. [14] presented a Laplacian bidirectional maximum margin criterion to avoid inverse problem. However, current manifold learning algorithms [6–14] might be unsuitable for pattern recognition tasks in that they concentrate on representing the high-dimensional data with low-dimensional data instead of classification or that they only considered the locality and could not give a clear nonlinear map when applied to a new sample, such as ISOMAP and LLE. Moreover, some of recent researches [15–21] presented a novel feature extraction method from multi-scale and sparse views.

Fortunately, Yan et al. [9] proposed a newly general framework called graph embedding for dimensionality reduction, from which many algorithms, such as PCA, LDA, LPP, ISOMAP, LLE, Laplacian Eigenmap can all be reformulated. Using the graph embedding framework as a platform, they developed a novel dimensionality reduction algorithm, marginal fisher analysis (MFA), to overcome the limitation of LDA. The powerful strength of their algorithm came from the intrinsic graph and the penalty graph that defined in local neighborhood. This suggested that their framework did not take the nonlocality into account. However, Yang et al. [10] proposed an unsupervised discriminant projection (UDP) algorithm considering the nonlocal and local quantities at the same time. The effect of neglecting nonlocality in MFA algorithm framework is that its projection direction cannot ensure that all the distances between two classes in low-dimension will be farer than that in observation space. This means some samples in different classes might be closer in feature space than in observation space. Therefore, this will degrade the recognition rates.

To address this disadvantage, we borrow the forms of UDP and MFA and propose a new criterion method with the cue of PCA and LDA: graph embedding discriminant analysis (GEDA). Firstly, the intrinsic graph is designed to characterize the intraclass compactness, the penalty graph formulated for interclass separability and the nonlocal graph out of intraclass for nonlocal quality. Then, based on these characterizations, we proposed a criterion that is similar to the classical fisher criterion. The optimal solutions can be obtained by solving a generalized eigenequation. GEDA not only has the same advantages of MFA compared to LDA mentioned in [9], such as no assumption on the data distributions, obtaining more projections and more separability of different classes, but also captures the sum scatter of nonlocal neighborhood with the form of nonlocal graph instead of global scatter that PCA captures. Furthermore, because we absorb the nonlocal property out of intraclass and use the maximal criterion, the solutions of GEDA are optimal in total and more robust than MFA and LDA in small sample size problems. In a word, the main contribution of our paper is threefold:

1. This paper proposes a novel graph embedding discriminant analysis (GEDA) method. GEDA seeks to find a set of perfect projections that not only can impact the samples of intraclass and maximize the margin of interclass, but also can maximize the nonlocal scatter at the same time.
2. This paper provides the theoretical foundations of the proposed method. Based on the above theory, three graphs can be constructed using the relatively low-dimensional data in PCA subspace and the singularity difficulty can be avoided without losing important information.
3. The proposed method is applied to face recognition and is examined on the Yale, ORL and AR face image databases. The experimental results show that GEDA consistently outperforms LDA and MFA when the training sample size per class is small.

The rest of this paper is organized as follows: we review the linear methods in Sects. 2, 3 develops the relevant theory and method of GEDA, experiments are shown in Sect. 4 and finally, Sect. 5 offers our conclusions.

2 Outline of linear methods

Let us consider a set of N samples $\{x_1, x_2, \ldots, x_N\}$ taking values in an m-dimensional space and assume that each sample belongs to one of c classes. Let us also consider a linear transformation mapping the original m-dimensional space into an n-dimensional feature space, where $m > n$. The new feature vectors $y_k \in R^n$ are defined by the following linear transformation:

$$y_k = A^T x_k (k = 1, \ldots, N) \tag{1}$$

where $A \in R^{m \times n}$ is a transformation matrix.

2.1 Linear discriminant analysis (LDA)

Linear discriminant analysis is a supervised learning algorithm. Let l denotes the total class number and l_i denotes the number of training samples in the ith class. Let x_i^j denotes the jth sample in ith class, $\bar{x}$ be the mean of all the training samples, $\bar{x}_i$ be the mean of the ith class. The between-class and within-class scatter matrices can be evaluated by the following:

$$S_b = \sum_{i=1}^{l} l_i (\bar{x}_i - \bar{x}) \, (\bar{x}_i - \bar{x})^T, \tag{2}$$

$$S_w = \sum_{i=1}^{l} \sum_{j=1}^{l_i} (x_i^j - \bar{x}_i)(x_i^j - \bar{x}_i)^T. \tag{3}$$

Linear discriminant analysis aims to find an optimal projection Φ_{opt} such that the ratios of the between-class scatter to within-class scatter is maximized, that is,

$$\Phi_{\text{opt}} = \arg\min_{\Phi} \frac{\left|\Phi^T S_b \Phi\right|}{\left|\Phi^T S_w^T \Phi\right|} = [\phi_1\ \phi_2\ \ldots\ \phi_n], \tag{4}$$

where $\{\phi_i | i = 1, 2, \ldots, n\}$ is the set of generalized eigenvectors of S_b and S_w corresponding to the n largest generalized eigenvalues $\{\lambda_i | i = 1, 2, \ldots, n\}$, that is,

$$S_b \phi_i = \lambda_i S_w^{-1} \phi_i, i = 1, 2, \ldots, n. \tag{5}$$

Note that there are at most $c-1$ nonzero generalized eigenvalues.

2.2 Principal component analysis (PCA)

A fundamental unsupervised dimensionality reduction method is PCA. Let S_T be the total scatter matrix:

$$S_T = \sum_{i=1}^{N} (x_i - \bar{x})\ (x_i - \bar{x})^T, \tag{6}$$

where $\bar{x}$ is the mean of all the training samples. The PCA transformation matrix is defined as

$$P_{\text{opt}} = \arg\max_{P} [tr(P^T S_T P)] = [p_1\ p_2 \ldots p_n], \tag{7}$$

where p_i $(i = 1, 2, \ldots, n)$ is the eigenvector corresponding to the largest eigenvalue of S_T.

2.3 Marginal fisher analysis (MFA)

For a classification problem, the sample set for model training is represented as a matrix $X = [x_1, x_2, \ldots, x_N]$, where $x_i \in R^m (i = 1, 2, \ldots, N)$, N is the sample number and m is the feature dimension. For supervised learning problems, the class label of the sample x_i is assumed to be $c_i \in \{1, 2, \ldots, N_c\}$, where N_c is the number of classes. Let π_c and n_c denote the index set and number of the samples belonging to the cth class, respectively.

Let $G = \{X, W\}$ be an undirected weighted graph with vertex set X and similarity matrix $W \in R^{N \times N}$. It is easy to see that W is a symmetric matrix and each element of W measures the similarity of a pair of vertices. The diagonal matrix D and the Laplacian matrix L of a graph G are defined as

$$L = D - W, D_{ii} = \sum_{i \neq j} W_{ij} \tag{8}$$

In MFA algorithm, intrinsic graph and penalty graph are introduced. Intraclass compactness is characterized from the intrinsic graph by the term

$$\begin{aligned} S_{\text{in}} &= \sum_i \sum_{i \in N_K^+(j)\ \text{or}\ j \in N_K^+(i)} \left\|\omega^T x_i - \omega^T x_j\right\|^2 \\ &= 2\omega^T X (D^{\text{in}} - W^{\text{in}}) X^T \omega \end{aligned} \tag{9}$$

$$\overset{\text{in}}{W}_{ij} = \begin{cases} 1, & \text{if } (i,j) \in \pi_c, \quad \text{and} \quad i \in N_K^+(j) \quad \text{or} \quad j \in N_K^+(i), \\ 0, & \text{else.} \end{cases}$$

where $N_K^+(i)$ indicates the index set of the K-nearest neighbors of the sample x_i in the same class. Then, we have $L^{\text{in}} = D^{\text{in}} - W^{\text{in}}$. From the form of S_{in}, it is easy to see that it exactly characterizes the sum scatter of the samples in K-neighborhood in the same class.

Interclass separability can be characterized by a penalty graph with the term

$$\begin{aligned} S_p &= \sum_i \sum_{i \in P_{K_p}(j)\ \text{or}\ j \in P_{K_p}(i)} \left\|\omega^T x_i - \omega^T x_j\right\|^2 \\ &= 2\omega^T X (D^p - W^p) X^T \omega \end{aligned} \tag{10}$$

$$\overset{p}{W}_{ij} = \begin{cases} 1, & \text{if } (i,j) \in P_{K_p}(c_i) \quad \text{or} \quad (i,j) \in P_{K_p}(c_j) \\ 0, & \text{else.} \end{cases}$$

where $P_{K_p}(c)$ is a set of data pairs that are in the K_p nearest pairs among the set $I = \{(i,j) | i \in \pi_c\ , j \notin \pi_c\}$. Then, we have $L^p = D^p - W^p$. From the form of S_p, it is easy to see that it exactly characterizes the sum scatter of the margin in K-neighborhood of interclass.With the intrinsic graph and penalty graph, MFA algorithm solves corresponding generalized eigen-equation of the following criterion:

$$\omega^* = \arg\min_{\omega} \frac{\omega^T X L^{\text{in}} X^T \omega}{\omega^T X L^p X^T \omega}. \tag{11}$$

3 Graph embedding discriminant analysis (GEDA)

3.1 Discussions and our motivation

Linear discriminant analysis searches for the directions that minimize the ratio between the intraclass and interclass scatters globally. But it cannot guarantee that any pair of classes will be more apart from each other in feature space because of not considering the local qualities. This means it may happen that two mutually distant classes (or samples) may be closer in feature space, that is, it cannot maximize the margin of interclass in locality because of not considering the locality. Therefore, LDA does not necessarily yield a perfect projection. But it does give us a good direction to develop our algorithm: the smaller the distance of intraclass is and the bigger the distance of interclass is, the higher the classifier accurate might be.

Principal component analysis seeks projections with maximal variances. In other words, when it points to classification, the bigger are the variances the better are the results from PCA's viewpoint. As an unsupervised learning algorithm, PCA only considers the global scatter information. This gives us a cue that the bigger the global scatter is the higher is the classifier rate even if it is not true

in some cases. This also suggests if the distance of each pair of classes or samples in different classes becomes farer in feature space, the higher rate might be obtained.

What is a perfect projection from the viewpoints combining PCA and LDA? In our opinions, for a classification problem, the perfect projections may be those that can impact the samples of intraclass, maximize the margin of interclass and the sum scatter of local K-neighborhood out of intraclass at the same time. However, none of the algorithms mentioned above can guarantee these three requests at the same time. Our motivation is to absorb and combine the ideas of PCA and LDA to develop an algorithm to find a set of projections that can simultaneously satisfy these three requests based on graph embedding, that is, GEDA.

3.2 Marginal graph and nonlocal graph of GEDA

In GEDA algorithm, there are three graphs: intrinsic graph, marginal graph and nonlocal graph. We call them as basic graphs. Usually, there are two variations for weighting the nearest neighbors: heat kernel and simple-minded [8]. In this paper, we adopt the form of "simple-minded".

For simplicity, the following graphs we construct only give the form of K-neighborhood (δ-neighborhood graph has the same form). Similar to 2.3., marginal separability can be characterized by the marginal graph defined with the term:

$$S_m = \sum_i \sum_j \overset{m}{W}_{ij} \left\| \omega^T x_i - \omega^T x_j \right\|^2 = 2\omega^T X (D^m - W^m) X^T \omega \tag{12}$$

$$\overset{m}{W}_{ij} = \begin{cases} 1, & \text{if } i \in N(j) \quad \text{or} \quad j \in N(i) \\ 0, & \text{else.} \end{cases}$$

where W denotes the similar matrix, and $N(i)$ indicates the index set of the K'-nearest neighbors of the sample x_i in the different class. Then, we have $L_m = D^m - W^m$. From the form of S_m, it is easy to see that it exactly characterizes the sum scatter of the samples in marginal graph.

The nonlocal graph is constructed as follows:

$$S_n = \sum_i \sum_j \overset{n}{W}_{ij} \left\| \omega^T x_i - \omega^T x_j \right\|^2 = 2\omega^T X (D^n - W^n) X^T \omega \tag{13}$$

$$\overset{n}{W}_{ij} = \begin{cases} 1, & \text{if } i \in N_k^-(j) \quad \text{or} \quad j \in N_k^-(i) \\ 0, & \text{else.} \end{cases}$$

where $N^-(i)$ indicates the index set out of the K-nearest neighbors of the sample x_i or in the K-nearest neighbors of the sample x_i but not in the same class. Then, we have $L^{\mathrm{n}} = D^{\mathrm{n}} - W^{\mathrm{n}}$. From the form of S_{n}, it is easy to see that it exactly characterizes the sum scatter of the samples in nonlocal graph.

3.3 The criterion of GEDA

Graph embedding discriminant analysis seeks to find a set of perfect projections. The projections not only can impact the samples of intraclass and maximize the margin of interclass, but also can maximize the nonlocal scatter out of intraclass at the same time. Based on the three basic graphs mentioned above, we have the criterion of GEDA:

$$\omega^* = \arg\max_{\omega} \left(\frac{\omega^T X L^m X^T \omega}{\omega^T X L^{\mathrm{in}} X^T \omega} + \frac{\omega^T X L^n X^T \omega}{\omega^T X L^{\mathrm{in}} X^T \omega} \right) \tag{14}$$

$$= \arg\max_{\omega} \frac{\omega^T X (L^m + L^n) X^T \omega}{\omega^T X L^{\mathrm{in}} X^T \omega} \tag{15}$$

The above criterion is formally similar to the Fisher criterion since they are both Rayleigh quotients. Therefore, we can obtain its optimal solutions by solving a generalized eigen-equation:

$$X(L^m + L^n) X^T \omega = \lambda X L^{\mathrm{in}} X^T \omega. \tag{16}$$

where λ is generalized eigenvalue, ω is generalized eigenvector correspondingly.

Linear discriminant analysis searches for the directions that minimize the ratio between the intraclass and interclass scatters. That is to say when the distance of intraclass is smaller and interclass bigger, the classifier accurate will be higher. PCA seeks projections with maximal variances. In other words, when it points to classification, the bigger are the variances the better the results in PCA's viewpoint. MFA finds the projections that minimize the ratio between the intraclass compactness and the interclass separability. UDP seeks the projections that maximize the local scatter and maximize the nonlocal scatter in the same time. Even if the authors did not point out clearly, MFA and UDP absorbed the idea of LDA and PCA in their frameworks, in which the ideas were shown in penalty graph and nonlocality, respectively, in a sense.

For a classification problem, the perfect projections may be those that can impact the samples of intraclass, maximize the margin of interclass and the sum scatter of local K-neighborhood out of intraclass at the same time. None of the algorithms mentioned above can guarantee these three requests at the same time. However, GEDA can achieve this propose. GEDA absorbs and develops the idea of the methods mentioned above in clear way: the first part of its criterion (14) impacts the samples of intraclass and maximizes the margin of interclass, and the second part maximizes the sum scatter nonlocal neighborhood at the same time. Therefore, we can say that GEDA finds a balance point of impacting the samples of intraclass, maximizing the margin of interclass and the sum scatter out of nonlocal

neighborhood at the same time. Thus, it outperforms the algorithms mentioned above.

3.4 GEDA algorithm for SSS problem

Usually, feature dimension in observation space is very high and the numbers of training samples in one class are undecided and small. On the one hand, it is very expensive computationally to construct the large size of three basic graphs directly with high-dimensional data. On the second hand, it should be guaranteed that $XL^{\text{in}}X^T$ is irregular. Therefore, it is helpful to preprocess with PCA for dimensionality reduction. Based on the following theory, three graphs can be constructed using the relatively low-dimensional data in PCA subspace and the singularity difficulty can be avoided without losing important information. The relevant theory is given below.

Suppose $\beta_1, \beta_2, \ldots, \beta_{m'}$ are m' orthonormal eigenvectors of S_T and the first n' eigenvectors corresponding to positive eigenvalues $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_{n'}$, where $n' = \text{rank}(S_T)$. Define the subspace $\Psi_T = \text{span}(\beta_1, \beta_2, \ldots, \beta_{n'})$ and denote its orthogonal complement $\Psi_T^{\perp} = \text{span}(\beta_{n'+1}, \beta_{n'+2}, \ldots, \beta_{m'})$. Obviously, Ψ_T is the range space of S_T and $\Psi_T^{\perp}$ is the corresponding null space.

Lemma 1 [12] *Suppose that A is ann × nnonnegative definite matrix and φ is an n-dimensional vector, then $\varphi^T A\varphi = 0$if and onlyif $A\varphi = 0$.*

Corollary 1 *If $S_T$is singular, $\varphi^T S_T \varphi = 0$if and only if $S_T\varphi = 0$. Since $R^{n'} = \text{span}(\beta_1, \beta_2, \ldots, \beta_{n'})$, for an arbitrary $\varphi \in R^{n'}$, φcan be denoted by*

$$\varphi = k_1\beta_1 + \cdots + k_{n'}\beta_{n'} + k_{n'+1}\beta_{n'+1} + \cdots + k_{m'}\beta_{m'}. \quad (17)$$

Let $w = k_1\beta_1 + \cdots + k_{n'}\beta_{n'}$and $u = k_{n'+1}\beta_{n'+1} + \cdots + k_{m'}\beta_{m'}$,then, from the definitionof$\Psi_T$and $\Psi_T^{\perp}$, φ can be denoted by $\varphi = w + u$, wherew $\in \Psi_T$and $u \in \Psi_T^{\perp}$.

Proposition 1 *The compression mapping $F : R^{m'} \to \Psi_T$ defined by $\varphi = w + u \to w$ is a linear transformation. Denote $J(w) = \frac{w^T X(L^m + L^n)X^T w}{w^T X L^{\text{in}} X^T w}$, and $\tilde{J}(\varphi) = \frac{\varphi^T \tilde{X}(L^m + L^n)\tilde{X}^T \varphi}{\varphi^T \tilde{X} L^{\text{in}} \tilde{X}^T \varphi}$, then we have the theorem:*

Theorem 1 *Under the compression mapping F, the GEDA criterion satisfies $J(w) = \tilde{J}(\varphi)$.*

Proof Let $Q = (\beta_1, \beta_2, \ldots, \beta_{n'})$. Then, we get $\tilde{X} = Q^T X$, that is $X = Q\tilde{X}$

$$J(w) = \frac{w^T X(L^m + L^n)X^T w}{w^T X L^{\text{in}} X^T w} = \frac{\omega^T (Q\tilde{X})(L^m + L^n)(Q\tilde{X})^T \omega}{\omega^T (Q\tilde{X}) L^{\text{in}} (Q\tilde{X})^T \omega} = \frac{\varphi^T \tilde{X}(L^m + L^n)\tilde{X}^T \varphi}{\varphi^T \tilde{X} L^{\text{in}} \tilde{X}^T \varphi} = \tilde{J}(\varphi),$$

where $\varphi = Q^T w$. □

According to theorem 1, we can draw a conclusion that the basic graphs can be constructed in PCA transformed space Ψ_T without losing any effective discriminant information with respect to the GEDA criterion. From linear algebra theory, compression mapping F is isomorphic to an n'-dimensional Euclidean space $R^{n'}$ and the corresponding isomorphic mapping is $w = Pv$, where $P = (\beta_1, \beta_2, \ldots, \beta_{n'})$, $v \in R^{n'}$, which is a one-to-one mapping from $R^{n'}$ onto Ψ_T.

By the property of isomorphic mapping and theorem 1, it is easy to see that the following theorem holds:

Theorem 2 *Let $w = Pv$ be an isomorphic mapping from $R^{n'}$ onto Ψ_T. Then $w^* = Pv^*$is the stationary point of the GEDA's criterion function $\tilde{J}(\varphi)$if and only if $v^*$is the stationary point of thefunction J(w).*

Proposition 2 *If $v_1, v_2, \ldots, v_d$ are the generalized eigenvectors of (16) corresponding to the d largest eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$, then $w_1 = Pv_1, \ldots, w_d = Pv_d$are the optimal projection axes of GEDA.*

It should be noted that the above derivation is based on the whole range space of S_T (i.e., all nonzero eigenvectors of S_T are used to form this subspace). In practice, however, we always choose the number of principal eigenvectors, m, smaller than the real rank of S_T such that most of the spectrum energy is retained and $\tilde{X}L^{\text{in}}\tilde{X}^T$ is well-conditioned (at least nonsingular) in the transformed space. In this case, the developed theory can be viewed as an approximate one and the generalized eigenvectors of $\tilde{J}(\varphi)$ can be calculated directly using the classical algorithm.

In summary of the preceding description, the following provides the GEDA algorithm:

Step 1. Perform PCA transform of data: calculate the k eigenvectors $\alpha_1, \alpha_2, \ldots, \alpha_k$ corresponding to k largest positive eigenvalues, let $P = (\alpha_1, \alpha_2, \ldots, \alpha_k)$. Then, we get $\tilde{X} = P^T X$.

Step 2. Construct the three basic graphs in PCA subspace and obtain L^m, L^n, L^{in}.

Step 3. Solve the generalized eigen-equation (16) and obtain the generalized eigenvectors $v_1, v_2, \ldots, v_d$ corresponding to the d largest positive eigenvalues. Then, the d projection axes of GEDA are $w_j = Pv_j, j = 1, \ldots, d$.

After obtaining the projection axes, we can form the following linear transform for a new sample x:

$$y = W^T x, \text{ where } W = (w_1, w_2, \ldots, w_d). \quad (18)$$

The feature vector y is used to represent the sample in the low-dimension feature space and used for recognition purposes.

Graph embedding discriminant analysis (GEDA) and MFA are both supervised subspace learning techniques

based on graph embedding. They are closely related to each other on the intrinsic graph in local neighborhood. Their criteria and idea, however, are quite different: firstly, GEDA adopts the maximal criterion and MFA adopts the minimal criterion instead. Secondly, GEDA introduces the nonlocal graph to characterize the nonlocal scatter, but MFA does not. Thus, GEDA will be more robust than MFA in data predicting. Thirdly, MFA finds the projections that only minimize the ratio between the intraclass compactness and the interclass separability, and thus, it cannot guarantee the three requests mentioned in Sect. 3.3. But EGDA can make it.

4 Experiments

To evaluate the proposed GEDA algorithm, we systematically compare it with the PCA, LDA and MFA algorithm in three face databases: ORL, Yale and AR. The ORL database is used to evaluate the performance of GEDA under conditions where the pose and sample size are varied. The Yale database is used to examine the system performance when both facial expressions and illumination are varied. The AR database is employed to test the performance of the system under conditions where there is a variation over time, in facial expressions and in lighting conditions. Euclidean distance and nearest neighborhood classifier are used in all the experiments. The number of local neighborhood, K, is chosen as $K = l-1$, where l denotes the number of training samples per class.

4.1 Experiments on ORL database

The ORL database is used to evaluate the performance of GEDA under conditions where the pose, face expression and sample size vary. The ORL face database contains images from 40 individuals, each providing 10 different images. The facial expressions and facial details (glasses or no glasses) also vary. The images were taken with a tolerance for some tilting and rotation of the face of up to 20°. Moreover, there is also some variation in the scale of up to about 10 %. All images normalized to a resolution of 56 × 46. Figure 1 shows sample images of one person. Note that LDA, MFA and GEDA all involve a PCA phase. Some projections of null space of S_b are used in LDA. The results are show in Table 1. From this experiment, we find that GEDA can achieve higher recognition rate on ORL face database when the training sample number is small. The reason is that the local neighbor relationship can proved important discriminant information. Moreover, the nonlocal graph and marginal graph of SDP can provide more discriminant information than the penalty graph cannot provide, and thus, the top recognition rates of GEDA is higher than MFA (Fig. 2).

4.2 Experiments on Yale database

The Yale face database contains 165 images of 15 individuals (each person providing 11 different images) under various facial expressions and lighting conditions. In our experiments, each image was manually cropped and resized to 100 × 80 pixels. Figure 3 shows sample images of one person. For computational effectiveness, we down sample it to 50 × 40 in this experiment. The experiment was performed using the first six images (i.e., center-light, with glasses, happy, left light, without glasses and normal) per class for training and the remaining five images (i.e., right light, sad, sleepy, surprised, and winking) for testing. For feature extraction, we used, respectively, PCA (eigenface), LDA (Fisherface), MFA and the proposed GEDA. Note that LDA, MFA and GEDA all involve a PCA phase. In this phase, we keep 90% image energy. The maximal recognition rate of each method and the corresponding dimension are given in Table 2. As is shown in Table 2 and Fig. 4, and the top recognition rate of GEDA is significantly higher than the other methods. Why can GEDA significantly outperform the other algorithm? An important reason may be that GEDA not only characterizes

Table 1 The maximal recognition rates (%) of PCA, LDA, MFA, GEDA on the ORL database and the corresponding dimensions (shown in parentheses) when the first 3, 4, 5 samples per class are used for training and the remaining for testing

Training sample number	PCA	LDA	MFA	GEDA
5	89.00 (34)	92.50 (33)	94.00 (50)	94.50 (60)
4	88.33 (73)	91.25 (38)	93.75 (40)	94.17 (38)
3	86.07 (95)	87.86 (41)	88.57 (40)	89.64 (35)

Fig. 1 The sample images of one person from ORL face database

the nonlocal scatter but also builds the adjacency relationship of data points using *K*-nearest neighbors at the same time, thus eliminates more negative influence of outliers.

4.3 Experiments on the AR face database

The AR face [40], [41] contains over 4,000 color face images of 126 people (70 men and 56 women), including frontal views of faces with different facial expressions, lighting conditions and occlusions. The pictures of 120 individuals (65 men and 55 women) were taken in two sessions (separated by 2 weeks), and each section contains 13 color images. Seven images of these 120 individuals are selected and used in our two experiments. The face portion of each image is manually cropped and then normalized to 50 × 40 pixels. The sample images of one person are shown in Fig. 5. These images vary as follows: (1) neutral expression (2) smiling (3) angry (4) screaming (5) left light

Table 2 The maximal recognition rates (%) of PCA, LDA, MFA, GEDA on the Yale database and the corresponding dimensions when the first six samples per class are used for training

	PCA	LDA	MFA	GEDA
Recognition rates (%)	92	93.33	94.67	97.33
Dimensions	30	18	16	24

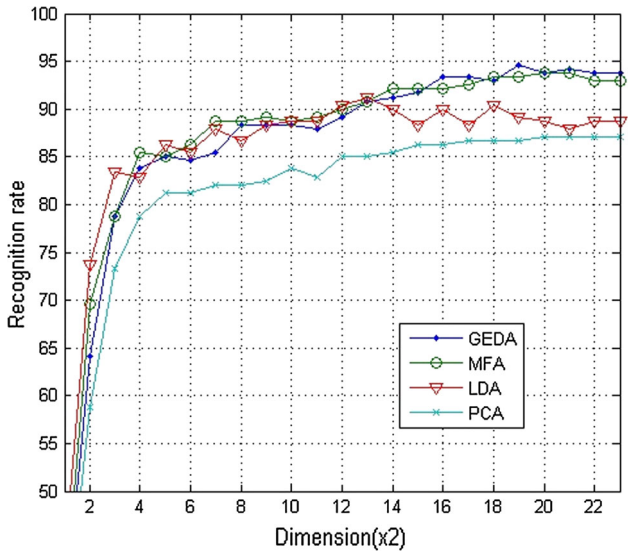

Fig. 2 The recognition rates (%) of PCA, LDA, MFA and GEDA versus the dimensions when the first four images per person were used for training on the ORL face database

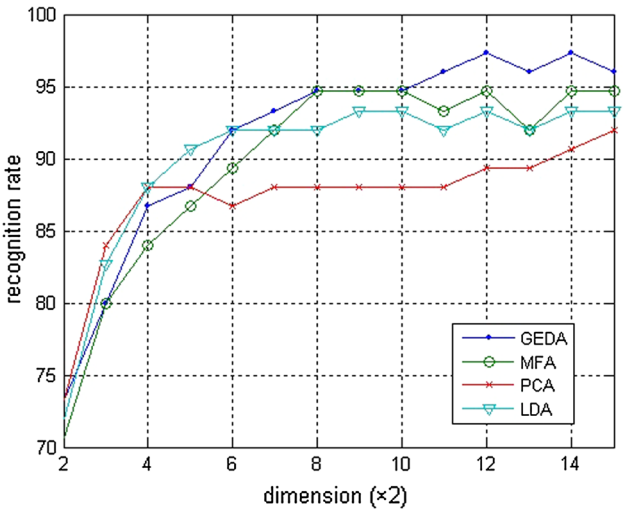

Fig. 4 The recognition rates (%) of PCA, LDA, MFA and GEDA versus the dimensions when the first six images per person were used for training on the Yale face database

Fig. 3 Sample images of one person in the Yale database

on (6) right light on (7) all sides light on. We select them to form two subsets: subset 1 and subset 2.

4.3.1 *Different facial expressions and lighting conditions but in the same section (time)*

In our first experiment, the first 7 images in the first section (the first line images in Fig. 5) are used as subset 1. The first l images (l varies from 3 to 5) in subset 1 are selected from the image gallery of each individual to form the training sample set. The remaining 7-l images are used for testing. For each l, PCA, LDA, MFA and GEDA are, respectively, used for face recognition. In the PCA phase of LDA, MFA and GEDA, the energy is set to be 95 or 96%. The K-nearest neighborhood parameter K is chosen as $K = l-1$. The dimension step is set to be 5. Finally, a nearest-neighbor classifier is employed for classification. The maximal recognition rate and the dimension are shown in Table 3. The recognition rate curves versus the variation of dimensions are shown in Figs. 6 and 7. From Table 3, Figs. 6 and 7, we can see first that GEDA significantly outperforms MFA and LDA, and second that as supervised methods, GEDA is more robust than MFA and LDA when there are different facial expressions and lighting conditions, irrespective of the variation in training sample size and dimensions. These two points are consistent with the experimental results in Sects. 4.1 and 4.2. Moreover, it should be noted that the recognition rate of MFA is lower than LDA and affected the most when the training numbers are 3 and 4. The reason may be that when there are no lighting samples in training set, the local neighbor of MFA cannot reflect or predict the relationship with the lighting samples in text set in a

Table 3 The maximal recognition rates (%) of PCA, LDA, MFA, GEDA on the subset 1 of AR database and the corresponding dimensions (shown in parentheses) when the first 3, 4, 5 samples per class are used for training and the remaining for testing

Training sample number	PCA	LDA	MFA	GEDA
5	75 (120)	95.5 (115)	97.08 (170)	98.33 (135)
4	66.67 (120)	91.33 (115)	88.06 (125)	93.33 (100)
3	73.54 (95)	94.17 (100)	93.75 (100)	94.79 (95)

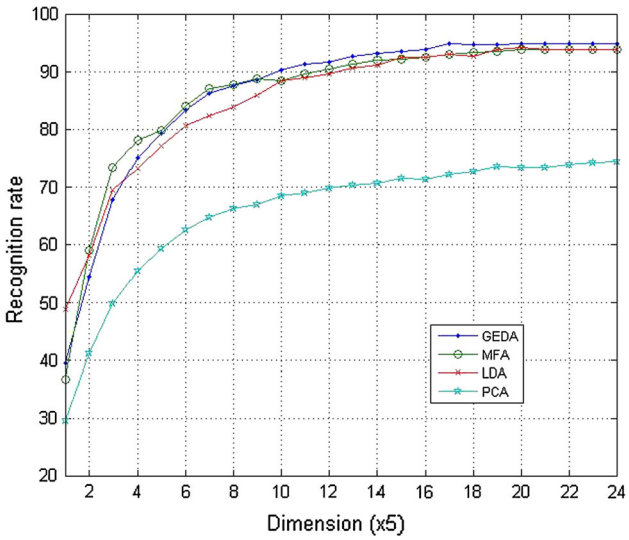

Fig. 6 The recognition rates (%) of PCA, LDA, MFA and GEDA versus the dimensions when the first three images per person were used for training and the remaining 4 images per person for testing on the subset 1 of AR face database

Fig. 5 Sample images of one subject of the AR database. The first line and the second line images were taken in different time (separated by 2 weeks)

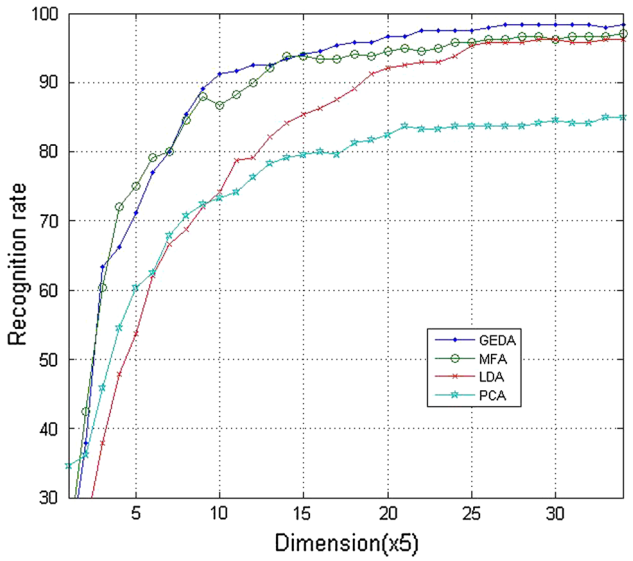

Fig. 7 The recognition rates (%) of PCA, LDA, MFA and GEDA versus the dimensions when the first five images per person were used for training and the remained two images per person for testing on the subset 1 of AR face database

certain sense. However, when the training number is 5, that is, there is one lighting image on the training set, and the local graph and penalty graph can characterize the property of the data. Thus, MFA is superior to LDA when the training number is 5.

4.3.2 *Not only with different facial expressions and lighting conditions but also in different section (time)*

In our second experiment, the first 3 images in the first section (the first line images in Fig. 5) and the 4th to 7th images in the second section (4th to 7th images in the second line in Fig. 5) are used as subset 2. There are 7 images for each subject. The first 3 images are selected from the image gallery of each individual to form the training sample set. The remaining 4 images are used for testing. The energy in PCA step is kept 95%. The top recognition rate and corresponding dimension are shown in Table 4. The top recognition rate of GEDA is the highest. This experiment also supports our analysis mentioned above and suggests that GEDA has more robust than MFA and LDA on facial expressions and lighting conditions and time variations.

Table 4 The maximal recognition rates (%) of PCA, LDA, MFA, GEDA on the subset 2 of AR database and the corresponding dimensions

	PCA	LDA	MFA	GEDA
Recognition rates (%)	39.58	52.08	50.20	53.54
Dimensions	110	110	100	100

4.4 Overall observations and evaluations of the experimental results

The above experiments show that the top recognition rate of GEDA is always higher than LDA and MFA. From the experiments, we can draw the following conclusions in details:

1. Graph embedding discriminant analysis consistently outperforms LDA, MFA and PCA despite the variation of dimensions and the number of training samples.
2. It should be pointed out that both MFA and GEDA can obtain more projections. However, when the numbers of projections are beyond the dimension number that LDA can provide, that is, rank of S_b, GEDA's projections are more effective than that of MFA (see Table 1 when training number is 5 and Fig. 7).
3. When the class number is not too large, such as in Yale and ORL database, LDA can achieve higher recognition rate in lower subspace. However, with the increase of dimensionality, LDA is significantly inferior to MFA and GEDA, which are shown in Figs. 2 and 4. On the other hand, if there are not or few lighting samples in training set, MFA is not necessary superior than LDA (see Figs. 6, 7; Table 3).
4. It should be noted that the recognition rates of GEDA often no less than MFA and LDA in the dimensions that MFA and LDA achieve their highest recognition rates (see Figs. 4, 6, 7).

5 Conclusion

In this paper, we develop a supervised discriminant technique, called graph embedding discriminant analysis (GEDA), for dimensionality reduction of high-dimensional data in small sample size cases. The projection of GEDA can be viewed as a linear approximation of the nonlinear map that uncovers and separates embeddings corresponding to different manifolds in the final embedding space. GEDA considers the local property, marginal property and nonlocal property and seeks to find a projection that not only can impact the samples of intraclass and maximize the margin of interclass, but also can maximize the nonlocal scatter at the same time. The consideration of the three aspects makes GEDA more intuitive and more powerful than LDA and MFA for classification tasks. Our experimental results on three popular face image databases demonstrate that GEDA is more effective than LDA and MFA in small sample size problems.

Acknowledgments The authors would like to thank the anonymous reviewers for their critical and constructive comments and suggestions. This work is partially supported by China Postdoctoral Science Foundation under grant No. 2011M500626, 2012M511479 and China National Natural Science Foundation under grant No. 61203247, 61273304, 61203376, 61202170, 61103067 and 61075056. It is also partially supported by The Project Supported by Fujian and Guangdong Natural Science Foundation under grant No. 2012J01281 and S2012040007289, respectively. It is also partially supported by the Fundamental Research Funds for the Central Universities.

References

1. Jain AK, Duin RPW, Mao J (2000) Statistical pattern recognition: a review. IEEE Trans Pattern Anal Mach Int 22(1):4–37
2. Joliffe I (1986) Principal component analysis. Springer, Berlin
3. Fukunnaga K (1991) Introduction to statistical pattern recognition, 2nd edn. Academic Press, New York
4. Martinez AM, Kak AC (2001) PCA versus LDA. IEEE Trans Pattern Anal Mach Int 23(2):228–233
5. He X, Niyogi P (2003) Locality preserving projections. In: Proceedings of 16th conference neural information processing systems
6. Tenenbaum JB, de Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290:2319–2323
7. Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290:2323–2326
8. Belkin M, Niyogi P (2003) Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput 15(6):1373–1396
9. Yan S, Xu D, Zhang B, Zhang H, Yang Q, Lin S (2007) Graph embedding and extensions: a general framework for dimensionality reduction. IEEE Trans Pattern Anal Mach Int 40–51
10. Yang J, Zhang D, Yang JY, Niu B (2007) Globally maximizing, locally minimizing: unsupervised discriminant projection with applications to face and palm biometrics. IEEE Trans Pattern Anal Mach Int 29(4):650–664
11. F. Chung (1997) Spectral Graph Theory. Reg Conf Ser Math no. 92
12. Golub GH, VanLoan CF (1996) Matrix computations, 3rd edn. Johns Hopkins University Press, US
13. Wan MH, Lai ZH, Shao J, Jin Z (2009) Two-dimensional local graph embedding discriminant analysis (2DLGEDA) with its application to face and palm biometrics. Neurocomputing 73: 193–203
14. Yang WK, Wang JG, Ren MW, Yang JY (2009) Feature extraction based on laplacian bidirectional maximum margin criterion. Pattern Recogn 42(11):2327–2334
15. Zhao CR, Liu CC, Lai ZH (2011) Multi-scale gist feature manifold for building recognition. Neurocomputing 74(17):2929–2940
16. Zhao CR, Lai ZH, Liu CC, Gu XJ, Qian JJ (2012) Fuzzy local maximal marginal embedding for feature extraction. Soft Comput 16(1):77–87
17. Miao DQ, Gao C, Zhang N, Zhang ZF (2011) Diverse reduct subspaces based co-training for partially labeled data. Int J Approx Reason 52(8):1103–1117
18. Yang WK, Sun CY, Zhang L (2011) A multi-manifold discriminant analysis method for image feature extraction. Pattern Recogn 44(8):1649–1657
19. Lai ZH, Wong WK, Jin Z, Yang J, Xu Y (2012) Sparse approximation to the eigensubspace for discrimination. IEEE Trans Neural Netw Learn Syst 23(12):1948–1960
20. Wan MH (2012) Maximum inter-class and marginal discriminant embedding (MIMDE) for feature extraction and classification. Neural Comput Appl 21(7):1737–1743
21. Wan MH, Yang GW, Jin Z (2011) Feature extraction based on Fuzzy local discriminant embedding (FLDE) with applications to face recognition. IET Comput Vision 5(5):301–308