Learning from Local and Global Discriminative Information for Semi-supervised Dimensionality Reduction

Mingbo Zhao, Haijun Zhang, Zhao Zhang

***Abstract*—Semi-supervised dimensionality reduction is an important research topic in many pattern recognition and machine learning applications. Among all the methods for semi-supervised dimensionality reduction, SDA and LapRLS are two popular ones. Though the two methods are actually the extensions of different supervised methods, we show in this paper that they can be unified into a regularized least square framework. However, the regularization term added to the framework focuses on smoothing only, it cannot fully utilize the underlying discriminative information which is vital for classification. In this paper, we propose a new effective semi-supervised dimensionality reduction method, called LLGDI, to solve the above problem. The proposed LLGDI method introduces a discriminative manifold regularization term by using the local discriminative information instead of only relying on neighborhood information. In this way, both the local geometrical and discriminative information of dataset can be preserved by the proposed LLGDI method. Theoretical analysis and extensive simulations show the effectiveness of our algorithm. The results in simulations demonstrate that our proposed algorithm can achieve great superiority compared with other existing methods.**

***Index Terms*—Dimensionality Reduction, Semi-supervised Learning, Local and Global Discriminative Information**

I. Introduction

Dealing with high-dimensional data has always been a major problem with the research of pattern recognition and machine learning. Typical applications of these include face recognition, document categorization, and image retrieval. Finding a low-dimensional representation of high-dimensional space, namely dimensionality reduction is thus of great practical importance. The goal of dimensionality reduction is to reduce the complexity of input space and embed high-dimensional space into a low-dimensional space while keeping most of the desired intrinsic information [1-2] [18-24]. Among all the dimensionality reduction techniques, Principle Compent Analysis (PCA) [3] and Linear Discriminant Analysis (LDA) [4] are two popular methods which have been widely used in many classification applications. PCA pursues the direction of maximum variance for optimal reconstruction. While LDA, as a supervised method, is to find the optimal projection V that maximizes the between-class scatter matrix S_b while minimizes the within-class scatter matrix S_w in the low-dimensional subspace. Due to the utilization of label information, LDA can achieve better classification results than those obtained by PCA if sufficient labeled samples are provided [4].

Though supervised methods generally outperform unsupervised methods, obtaining sufficient number of labeled samples for training can be problematic because labeling large number of samples is time-consuming and costly. On the other hand, unlabeled samples may be abundant and can easily be obtained in the real world. Thus, using semi-supervised learning methods [6-9], which incorporate both labeled and unlabeled samples into learning procedure, has become an effective option instead of only relying on supervised learning. Two well-known semi-supervised learning methods are GFHF [6] and LLGC [7]. These methods work in a tranductive way by propagating the label information from labeled set to unlabeled set via label propagation. But they cannot predict the class labels of new-coming samples hence suffering the out-of-sample problem. In contrast, semi-supervised dimensionality reduction methods not only reduce the dimensionality but also naturally solve the out-of-sample problem, which is more practical in real-world applications.

Many semi-supervised dimensionality reduction have been proposed during the past decade. Two widely-used methods are Semi-supervised Discriminant Analysis (SDA) [8] and Laplacian Regularized Least Square (LapRLS) [9]. These methods share the same concept for dimensionality reduction, i.e. they first construct the graph Laplacian matrix to approximate the manifold structure by using both labeled and unlabeled samples. They then perform dimensionality reduction by adding the graph Laplacian matrix as a regularized term to the original objective function of LDA and Regularized Least Square (RLS), respectively. Hence both the discriminative structure embedded in the labeled samples and

Mingbo Zhao and Zhao Zhang are with the Department of Electronics Engineering, City University of Hong Kong, Kowloon, Hong Kong S.A.R. (Email: mzhao4@cityu.edu.hk, cszzhang@gmail.com; Phone: +852 9517 9589)

Haijun Zhang is with the Shenzhen Key Laboratory of Internet Information Collaboration and the Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, P.R. China (Email: aarhzhang@gmail.com; Phone: +86 755 2603 3086)

the geometrical structure embedded in both labeled and unlabeled set can be preserved. Though they are initially based on different supervised methods, we show in this paper that SDA can be addressed equivalently into a regularized least square framework. Then, both SDA and LapRLS can be covered into a unified framework.

The connection between SDA and LapRLS helps us well understand their relationship. In fact, both SDA and LapRLS can preserve the manifold smoothness embedded in both labeled and unlabeled set, but LapRLS is to fix a linear model to the labels of labeled set. Since it is essentially derived from regression problem instead of classification problem, LapRLS may not preserve more underlying discriminative information in the labeled set; On the other hand, under the aforementioned least square framework, SDA can be addressed by fixing a linear model to the between-class scatter matrix. But it can only preserve the global discriminative information, while neglects the local discriminative information.

In this paper, we propose a new effective semi-supervised dimensionality reduction method, called LLGDI, which can preserve more local discriminative information. In our method, we introduce a discriminative manifold regularization term by using the local discriminative information instead of only relying on neighborhood information. In this way, both the local geometrical and discriminative information of dataset can be preserved by the proposed LLGDI method. The main contributions of this paper are summarized as follows:

1) We address SDA into a least square framework and establish the connections between SDA and LapRLS.

2) By analyzing the connections between SDA and LapRLS, we propose a new method, called LLGDI, which can preserve both the local geometrical and discriminative information of dataset and overcome the shortcomings of SDA and LapRLS.

This paper is organized as follows: In Section 2, we will present an effective label propagation procedure. In Section 3, we will introduce our soft label based linear discriminant analysis (SL-LDA) for semi-supervised dimensionality reduction. We will also build a close relationship between SL-LDA and W-LS in this section and propose a more efficient approach for solving SL-LDA. The simulation results are shown in Section 4 and the final conclusions are drawn in Section 5.

II. Brief Review of the Prior Work

A. Linear Discriminant Analysis (LDA)

The goal of LDA is to seek an optimal projection matrix $V^* \in R^{D\times d}$ that maximizes between-class scatter matrix while minimizes with-class scatter matrix. Suppose we have a set of l samples $X_l = \{x_1, x_2, \ldots, x_l\} \in R^{D\times l}$ belonging to c classes. Each sample is associated with a class c_i from $\{1,2,\ldots,c\}$. Denote $Y = \{y_1, y_2, \ldots, y_j\} \in R^{c\times l}$ as the class matrix, where $y_{ij} = 1$, if x_j belongs to the *ith* class; $y_{ij} = 0$, otherwise. We also denote $G = \{g_1, g_2, \ldots, g_j\} = \left(YY^T\right)^{-1/2} Y \in R^{c\times l}$ as the scaled class matrix, where $g_{ij} = 1/\sqrt{l_i}$, if x_j belongs to the *ith* class; $g_{ij} = 0$, otherwise. Since YY^T is diagonal matrix, it follows $GG^T = \left(YY^T\right)^{-1/2} YY^T \left(YY^T\right)^{-1/2} = I$ [14]. Then, Assuming the data matrix X_l are centered, the total-class, between-class and within-class scatter matrix S_t, S_b, S_w can be defined as

$$\begin{aligned} S_t &= \sum_{i=1}^{c} \sum_{x\in C_i} (x-\mu)(x-\mu)^T = X_l X_l^T \\ S_b &= \sum_{i=1}^{c} l_i (\mu_i - \mu)(\mu_i - \mu)^T = X_l G^T G X_l^T \\ S_w &= \sum_{i=1}^{c} \sum_{x\in C_i} (x-\mu_i)(x-\mu_i)^T = X_l X_l^T - X_l G^T G X_l^T \end{aligned} \quad (1)$$

where l_i is the number of samples in the *ith* class, μ_i is the mean of samples in the *ith* class, and μ is the mean of all labeled samples. Since $S_t = S_w + S_b$, the goal of LDA is equivalent to solving the problem as:

$$J(V) = \max Tr\left(\left(V^T (S_t + \lambda I) V \right)^{-1} V^T S_b V \right). \quad (2)$$

where λI is a multiply of identify matrix added as a regularized term for avoiding the singularity of S_t [5]. The optimal projection matrix V^* are formed by eigenvectors corresponding to the d largest eigenvalues of $(S_t + \lambda I)^{-1} S_b$.

B. Semi-supervised Discriminant Analysis (SDA)

SDA extends the conventional LDA to preserve the geometric structure by adding a manifold regularized term to the objective function of LDA. Let $X = \{X_l, X_u\} = \{x_1, x_2, \ldots, x_{l+u}\} \in R^{D\times(l+u)}$ be the data matrix where the first l and the remaining u columns are the labeled and unlabeled samples, respectively. The objective function of SDA can be given by:

$$J(V) = \max Tr\left(\left(V^T \left(S_t + \lambda_t I + \lambda_m X L X^T \right) V \right)^{-1} V^T S_b V \right). \quad (3)$$

where $L = D - W$ is the graph Laplacian matrix associated with both labeled and unlabeled set [10], W is the weight matrix defined as: $w_{ij} = 1$, if x_i is within the k nearest neighbor of x_j or x_j is within the k nearest neighbor of x_i; $w_{ij} = 0$, otherwise, D is a diagonal matrix satisfying $D_{ii} = \sum_{j=1}^{l+u} w_{ij}$, λ_m and λ_t are the two parameters balance the tradeoff between two terms. The optimal solution of SDA is then formed by eigenvectors of $\left(S_t + \lambda_t I + \lambda_m X L X^T\right)^{-1} S_b$.

Note that in LDA and SDA, their optimal solutions are not unique [12]. Let V^* be the optimal solution of SDA (LDA is a special case of SDA by setting $\lambda_m = 0$), then, $V^*\Xi$ is also the solution of SDA satisfying $\left(S_t + \lambda I + \lambda_m X L X^T\right)^{-1} S_b V = V\Lambda$, where $\Xi \in R^{d\times d}$ is an arbitrary diagonal matrix and Λ is the eigenvalue matrix. Hence in order to make the solution unique, a typical constraint can be imposed to the objective function of SDA as:

$$V^T \left(S_t + \lambda_t I + \lambda_m X L X^T \right) V = I \quad (4)$$

In this paper, we concentrate on the solution of SDA with or without this constraint and investigate the relationship with least square. We show in the following section that SDA can be

analyzed under a least square framework, and based on this view, we establish the equivalence between least square view of SDA and Lap-RLS.

III. A LEAST SQUARE VIEW OF SEMI-SUPERVISED DIMENSIONALITY REDUCTION

A. Least Square Semi-supervised Discriminant Analysis

Least square is another popular technique which has been widely used for regression and classification [14]. Let $T \in R^{c\times l}$ be a certain class indicator matrix and $b \in R^{1\times c}$ be the bias term, the goal of least square is to fix a linear model $t_j = V^T x_j + b^T$ by regressing X_l on T. It has been investigated that there is certain relationship between LDA and least square [11-13]. In this section, we further extend this relationship to the semi-supervised version and analyze SDA under a least square view. Specifically, let us consider a least square problem with both manifold term and Tikhonov term regularized (we refer it as LS-SDA):

$$J(V,b) = \min \sum_{j=1}^{l} \left\| V^T x_j + b^T - g_j \right\|_F^2 + \lambda_t \|V\|_F^2 + \lambda_m Tr\left(V^T X L X^T V\right) \quad (5)$$

where $G = \{g_1, g_2, \ldots, g_j\}$ is the scaled class matrix defined as above. For convenience, we rewrite Eq. (5) in a matrix form as:

$$J(V,b) = \min Tr\left(V^T X + b^T e - G\right) U \left(V^T X + b^T e - G\right)^T + \lambda_t Tr\left(V^T V\right) + + \lambda_m Tr\left(V^T X L X^T V\right) \quad (6)$$

where $U \in R^{(l+u)\times(l+u)}$ is a diagonal matrix with the first l and the remaining u diagonal elements as 1 and 0, respectively, $e \in R^{1\times(l+u)}$ is a unit vector with size $l+u$. By setting the derivative w.r.t. V and b to zero, we have:

$$\begin{cases} \left(XUX^T + \lambda_t I + \lambda_m X L X^T\right) V = XUT^T - XUe^T b \\ \left(eUe^T\right) b = eUG^T - eUX^T V \end{cases}, \quad (7)$$

Following Eq. (7), we can calculate the optimal projection matrix of Eq. (6) as:

$$V^* = \left(XL_t X + \lambda_t I + \lambda_m X L X^T\right)^{-1} X L_t G^T \quad (8)$$

where $L_t = U - Ue^T eU / eUe^T$ is used for centering the labeled samples, and following Eq. (1), we have $XL_t X^T = S_t$ and $XL_t G^T G L_t X^T = S_b$. Here, if we further define $H_b = XL_t G^T$, the optimal solution of Eq. (6) can be rewritten as $V^* = \left(S_t + \lambda_m X L X^T + \lambda_t I\right)^{-1} H_b$.

B. Equivalence between LS-SDA and SDA

We next build the equivalence between LS-SDA and SDA. This equivalence is based on the following theorem:

Theorem 1 [17]: *Given two matrix A and B, then AB and BA have the same non-zero eigenvalues. For each nonzero eigenvalue of AB, if the corresponding eigenvector of AB is v, then the corresponding eigenvector of BA is $u = Bv$.*

Recall that the solution of SDA is formed by the eigenvectors of $\left(S_t + \lambda_m X L X^T + \lambda_t I\right)^{-1} S_b = \left(S_t + \lambda_m X L X^T + \lambda_t I\right)^{-1} H_b H_b^T$. Based on Theorem 1, it has the same nonzero eigenvalues to the auxiliary matrix M:

$$M = H_b^T \left(S_t + \lambda_m X L X^T + \lambda_t I\right)^{-1} H_b = U \Sigma U^T \in R^{c\times c}. \quad (9)$$

where $U\Sigma U^T$ is the Singular Value Decomposition of M. According to Theorem 1 again, if U is the eigenvectors to the nonzero eigenvalues of M, $\left(S_t + \lambda_m X L X^T + \lambda_t I\right)^{-1} H_b U$ is its eigenvectors. Note that $\left(S_t + \lambda_m X L X^T + \lambda_t I\right) H_b$ is the solution of LS-SDA, we have $V_{SDA}^* = V_{LS-SDA}^* U \Xi$, where $\Xi \in R^{d\times d}$ is any diagonal matrix. We thus develop two SDA methods by choosing different Ξ. Let $\Xi = \Sigma^{-1/2}$ and $V_{SDA1}^* = V_{LS-SDA}^* U \Sigma^{-1/2}$, we have

$$V_{SDA1}^{*T} \left(S_t + \lambda_m X L X^T + \lambda_t I\right) V_{SDA1}^* = I. \quad (10)$$

which indicates that it is the optimal solution of SDA with the constraint $V^T \left(S_t + \lambda_m X L X^T + \lambda_t I\right) V = I$. For convenience, we referr it as SDA1; let $\Xi = I$ and $V_{SDA2}^* = V_{LS-SDA}^* U$, we have the following theorem:

Theorem 2: *Given the relationship, i.e. $V_{SDA2}^* = V_{LS-SDA}^* U$, we have $V_{SDA2}^* V_{SDA2}^{*T} = V_{LS-SDA}^* V_{LS-SDA}^{*T}$. Accordingly, for any two samples x_i and x_j, it follows:*

$$\left\| V_{SDA2}^{*T} \left(x_i - x_j\right) \right\|_F^2 = \left\| V_{LS-SDA}^{*T} \left(x_i - x_j\right) \right\|_F^2, \quad (11)$$

where $\left\| V^T \left(x_i - x_j\right) \right\|_F^2 = \left(x_i - x_j\right)^T V V^T \left(x_i - x_j\right)$ represents the Mahalanobis distance between x_i and x_j.

The proof of Theorem 2 is in appendix A. For convenience, we refer it as SDA2. Following Theorem 2, it indicates that when applying a distance-based classifier (such as kNN classifier), both LS-SDA and SDA2 can achieve the same classification results. Hence in practice, since solving LS-SDA can be more efficient as it is based on a least square framework, we can first calculate V_{LS-SDA}^*, then let $V_{LS-SDA}^* \to V_{SDA2}^*$. For SDA1, we need to further perform the SVD of M of $M = U\Sigma U^T$, then let $V_{LS-SDA}^* U \Sigma^{-1/2} \to V_{SDA1}^*$.

C. Equivalence between LS-SDA and Lap-RLS

Similar to LS-SDA, Lap-RLS extends the least square to its semi-supervised version by adding a manifold regularized term. The goal of Lap-RLS is to fix a linear model $y_j = V^T x_j + b^T$ by regressing X on Y and simultaneously to preserve the manifold smoothness embedded in both labeled and unlabeled set. The objective function and optimal solution of Lap-RLS can be given as

$$J(V,b) = \min \sum_{j=1}^{l} \left\| V^T x_j + b^T - y_j \right\|_F^2 + \lambda_t \|V\|_F^2 + \lambda_m Tr\left(V^T X L X^T V\right) \quad (12)$$

$$V^* = \left(S_t + \lambda_t I + \lambda_m X L X^T\right)^{-1} X L_t Y^T$$

We next show that given a certain condition, the optimal solution obtained by LapRLS is equivalent to that of LS-SDA. Actually, we have the following theorem:

Theorem 3: *given the condition that all classes in labeled set have the same number of samples, i.e.* $l_1 = l_2 = \cdots = l_c = n$, *the optimal solution of LS-SDA is equivalent to that of LapRLS.*

Proof of Theorem 3: Since the condition $l_1 = l_2 = \cdots = l_c = n$ holds, we have $1/\sqrt{l_1} = 1/\sqrt{l_2} = \cdots = 1/\sqrt{l_c} = 1/\sqrt{n}$. It then follows $G = 1/\sqrt{n}\,Y$ and we can rewrite the optimal solution of LS-SDA in Eq. (8) as $V^*_{LS-SDA} = 1/\sqrt{n}\left(S_t + \lambda_m XLX^T + \lambda_t I\right)^{-1} XL_tY^T$. Here, if we neglect the constant $1/\sqrt{n}$, we can observe that the optimal solution of LS-SDA is equival to that of LapRLS as in Eq. (12).

IV. Learning from Local and Global Discriminative Information

A. Motivation

The equivalence between LS-SDA and Lap-RLS as analyzed above help us well understand their connections. In fact, both two methods can preserve the manifold structure embedded in both labeled and unlabeled set. Lap-RLS is to fix a linear model to the class matrix. Since it is essentially derived from regression problem instead of classification problem, it may not preserve more underlying discriminative information embedded in both labeled and unlabeled set; On the other hand, by fixing a linear model to the between-class scatter matrix, LS-SDA is able to preserve the global discriminative information, but it neglects the local discriminative information. Hence to solve this problem, we present a new semi-supervised method in this section, which arms to preserve more discriminative information embedding in both labeled and unlabeled set [16].

B. Integration of local discriminative information

Let $N_k(x_j)$ be the k neighborhood set of x_j including itself, we denote $X_j = \{x_{j_0}, x_{j_1}, \ldots, x_{j_k}\} \in R^{D\times k}$ as the local data matrix formed by all samples in $N_k(x_j)$, where $\{j_1, j_1, \ldots, j_k\}$ is the index set of $N_k(x_j)$ and $j_1 = j$, $x_{j_1} = x_j$. We also denote $\widehat{G}_j = \{\hat{g}_{j_1}, \hat{g}_{j_2}, \ldots, \hat{g}_{j_k}\} \in R^{c\times k}$ as the local scaled class matrix in $N_k(x_j)$. More specifically, let $\hat{G} = \{\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_{l+u}\} \in R^{c\times(l+u)}$ be the global scaled class matrix of both labeled and unlabeled samples, $\widehat{G}_j$ can be viewed as a selection from $\hat{G}$ as:

$$\widehat{G}_j = \hat{G}S_j \tag{13}$$

where $S_j \in R^{(l+u)\times k}$ is the selected matrix with each element satisfying $(S_j)_{pq} = 1$, if $p = i_q$; $(S_j)_{pq} = 0$, otherwise. Then, assuming the samples in X_j are centered, the local between-class and total-class scatter matrixes can be constructed as $S_b^j = X_j\widehat{G}_j^T\widehat{G}_jX_j^T$, $S_t^j = X_jX_j^T$. The optimal local projection matrix and local scaled class matrix can be calculated simultaneously by maximizing the following objective function:

$$J_j\left(\widehat{V}_j, \widehat{G}_j\right) = \max_{\widehat{G}_j\widehat{G}_j^T = I} Tr\left(\left(\widehat{V}_j^T\left(X_jX_j^T + \lambda I\right)\widehat{V}_j\right)^{-1}\widehat{V}_j^T X_j\widehat{G}_j^T\widehat{G}_jX_j^T\widehat{V}_j\right) \tag{14}$$

Theorem 4: *Let* $J(V)$ *be the objective function of LDA as Eq. (2), then, we have*

$$\begin{aligned} J(V) &= \max Tr\left(\left(XX^T + \lambda I\right)^{-1} XG^TGX^T\right) \\ &= \max Tr\left(GX^T\left(XX^T + \lambda I\right)^{-1} XG^T\right). \\ &= \min Tr\left(G\left(X^TX + \lambda I\right)^{-1} G^T\right) \end{aligned} \tag{15}$$

The proof of Theorem 4 can be seen in Appendix B. Following Theorem 4, by simply performing notation substitutions to Eq. (15), i.e. $V_j \to V$ and $\widehat{G}_j \to G$, we can rewrite Eq. (14) as:

$$J_j\left(\widehat{V}_j, \widehat{G}_j\right) = J_j\left(\widehat{G}_j\right) = \min Tr\left(\widehat{G}_j\left(X^TX + \lambda I\right)^{-1}\widehat{G}_j^T\right) \tag{16}$$

However, Eq. (16) only gives a local objective function for calculating the optimal $\widehat{G}_j$. In order to calculate the global scaled class matrix $\hat{G}$, we sum the local objective function of Eq. (16) over all local patches, which can be formulated as:

$$\begin{aligned} J\left(\hat{G}\right) &= \min \textstyle\sum_{j=1}^{l+u} Tr\left(\widehat{G}_j\left(X^TX + \lambda I\right)^{-1}\widehat{G}_j^T\right) \\ &= \min Tr\left(\hat{G}\left(\textstyle\sum_{j=1}^{l+u} S_j\widehat{L}_jS_j^T\right)\hat{G}^T\right) = \min Tr\left(\hat{G}\hat{L}\hat{G}^T\right) \end{aligned} \tag{17}$$

where $\widehat{L}_j = \left(X_j^TX_j + \lambda I\right)^{-1}$ and $\hat{L} = \sum_{j=1}^{l+u} S_j\widehat{L}_jS_j^T$. In addition, we hope $\hat{G}$ can fix to its initial scaled class matrix G. Then, $\hat{G}$ can be calculated by minimizing the following objective function as:

$$J\left(\hat{G}\right) = \min \textstyle\sum_{j=1}^{l}\left\|\widehat{g}_j - g_j\right\|_F^2 + \lambda_m Tr\left(\hat{G}\hat{L}\hat{G}^T\right) \tag{18}$$

Thus, by setting the derivative w.r.t. $\hat{G}$ to zero, the optimal solution to the problem in Eq. (18) is $\hat{G} = \left(U + \lambda_m\hat{L}\right)^{-1}G$.

C. Subspace Learning for Dimensionality Reduction

Note that in the above subsection, we obtain the optimal global scaled class matrix $\hat{G}$ by integrating both local and global discriminative information. Then, assuming data matrix X is centered, the global between-class and total-class scatter matrixes can be constructed as $S_b = X\hat{G}^T\hat{G}X^T$, $S_t = XX^T$. The optimal global projection matrix $\hat{V}$ can be calculated by maximizing the following objective function as:

$$J\left(\hat{V}\right) = \max Tr\left(\left(\hat{V}^T\left(XX^T + \lambda I\right)\hat{V}\right)^{-1}\hat{V}^TX\hat{G}^T\hat{G}X^T\hat{V}\right). \tag{19}$$

Theorem 5: *Let* $J(V)$ *be the objective function of LDA as Eq. (2),* $L(V,b)$ *be the objective function of the following least square problem:*

$$L(V,b) = \min\left\|V^TX + b^Te - G\right\|_F^2 + \lambda\|V\|_F^2 \tag{20}$$

Then it follows $J(V) = L(V,b)$.

Proof of Theorem 5: By setting the derivatives w.r.t. V and b to zero, we have

$$\begin{cases} b = \left(eG^T - eX^TV\right)/ee^T \\ V = \left(XL_cX^T + \lambda I\right)^{-1}XL_cG^T \end{cases}, \tag{21}$$

where $e \in R^{1\times(l+u)}$ is a unit vector and $L_c = I - e^T e / ee^T$ is used for centering the samples by subtracting the mean of all samples. Hence with b and V in Eq. (21), the regression function $V^T X + b^T e - G$ can be written as:

$$V^T X + b^T e - G = G\left(L_c X^T \left(XL_c X^T + \lambda I\right)^{-1} XL_c - L_c\right) = G\left(N - L_c\right) \tag{22}$$

Where $N = L_c X^T \left(XL_c X^T + \lambda I\right)^{-1} XL_c$. By replacing V and $V^T X + b^T e - G$ in Eq. (20) with Eq. (21) and Eq. (22), we have:

$$\left\|V^T X + b^T e - G\right\|_F^2 + \lambda \|V\|_F^2 = Tr\left(GL_c \left(L_c X^T XL_c + \lambda I\right)^{-1} L_c G^T\right) \tag{23}$$

The derivation of Eq. (23) is shown in Appendix D. Note that X is centered, i.e. $X = XL_c$, then, following Eq. (23) and (15), we have $J(V) = L(V,b)$.

From Theorem 5, it shows that the cost error obtained by LDA is equivalent to that obtained by least square problem. Then, by simply performing notation substitution, i.e. $\hat{V} \to V$, $\hat{G} \to G$ in Eq. (21), we can rewrite Eq. (19) as:

$$J\left(\hat{V}\right) = J\left(\hat{V},\hat{b}\right) = \min \left\|\hat{V}^T X + \hat{b}^T e - \hat{G}\right\|_F^2 + \lambda \left\|\hat{V}\right\|_F^2 \tag{24}$$

In this paper, we propose our method by incorporating this linearity regularization $J\left(\hat{V},\hat{b}\right)$ into Eq. (18), in which we calculate the global scaled class matrix, global projection matrix and bias term simultaneously by minimizing the following objective function:

$$\begin{aligned} J\left(\hat{G},\hat{V},\hat{b}\right) &= \min Tr\left(\hat{G} - G\right)U\left(\hat{G} - G\right)^T \\ &+ \lambda_m Tr\left(\hat{G}\hat{L}\hat{G}^T\right) + \lambda_r \left(\left\|\hat{V}^T X + \hat{b}^T e - \hat{G}\right\|_F^2 + \eta \left\|\hat{V}\right\|_F^2\right) \end{aligned} \tag{25}$$

Since our method arms to learn local and global discriminative information, we refer it as LLGDI. We next show how to calculate the optimal solution in Eq. (18). By replacing $\left\|\hat{V}^T X + \hat{b}^T e - \hat{G}\right\|_F^2 + \eta \left\|\hat{V}\right\|_F^2$ in Eq. (25) with Eq. (24) and setting the derivative w.r.t. $\hat{G}$ to zero, we have

$$\hat{G} = GU\left(U + \lambda_m \hat{L} + \lambda_r L_c \left(L_c X^T XL_c + \eta I\right)^{-1} L_c\right)^{-1} . \tag{26}$$

Then, by replacing $\hat{G}$ in Eq. (21) with Eq. (26), we can obtain the optimal projection matrix $\hat{V}$ and bias term $\hat{b}$. The basic steps of the proposed LLGDI method can be shown in Table 1.

Table 1 Algorithm

Input:

Data matrix $X \in R^{D\times(l+u)}$, reduced matrix d and other related parameters.

Output:

The projection matrix $\hat{V} \in R^{D\times d}$.

Algorithm:

1. Construct the scaled class matrix G and the graph Laplacian matrix $\hat{L}$ as in Eq. (17).
2. Calculate the predicted scaled class matrix $\hat{G}$ as in Eq. (26).
3. Calculate the optimal projection matrix $\hat{V}$ by Eq. (21).
4. Output $\hat{V}$.

V. SIMULATIONS

A. Toy Examples for Synthetic Datasets

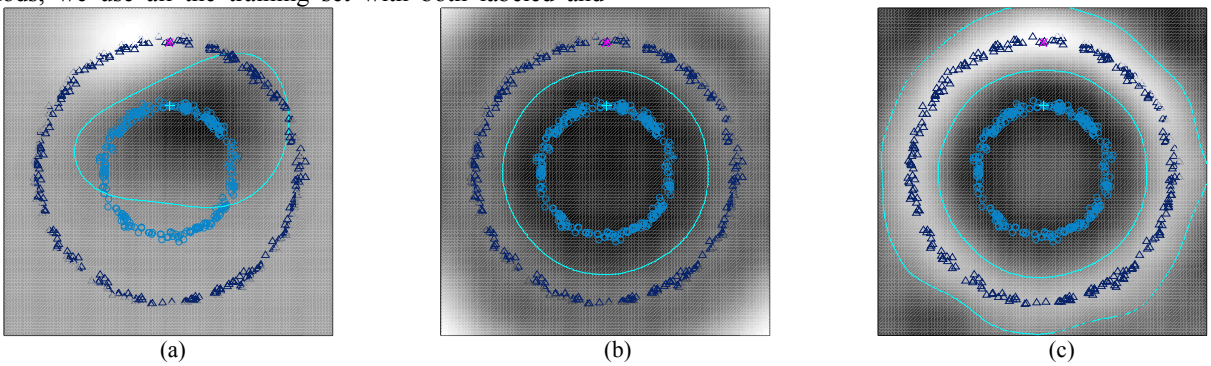
In this toy example, we generate a dataset with two classes, each follows a cycle distribution with the same core but different radius.. In each class, one sample is selected as labeled set and the remaining as unlabeled set. Since the distribution of two-cycle dataset is nonlinear, to handle this problem, we first perform KPCA to the two-moon dataset; we then use the output in the full-rank KPCA to train the linear methods [14]. Fig. 1 shows the gray images of decision surfaces and boundaries obtained by LDA, SDA and LLGDI. The gray value of each pixel represents the difference of distance from the pixel to its nearest labeled samples in different classes after dimensionality reduction. The decision boundaries are then formed by the pixels with the values equal to 0. In this example, we set the reduced dimensionality as 1. From Fig. 1 we can observe that for the two-cycle dataset, the decision boundary learned by LDA cannot classify the two classes. This indicates that given insufficient labeled samples, LDA fails to find the precise boundary between different classes. In contrast, by using the unlabeled samples to construct the manifold term for preserving the geometrical structure embedded into the dataset, SDA can find the precise decision boundary. In addition, the proposed LLGDI can achieve the best performance, as the decision boundary learned by LLGDI is more precise than those obtained by SDA. The improvement is reliable due to the fact that LLGDI preserves both local and global discriminative information embedded in dataset.

B. Classification

For classification problem, we use 8 real-world datasets to evaluate the performance of methods. The detailed information of dataset can be shown in Table 2. For each dataset, we randomly select l samples from each class as labeled set and u samples as unlabeled set. The test set is then formed by the selected or remaining samples. The data partitioning for each dataset is also given in Table 2.

Next, we compare our method with other supervised and semi-supervised dimensionality reduction methods. These methods include RLDA [5], RLS [14], SDA1 [8], SDA2, LS-SDA,Lap-RLS [9] and FME [15]. The simulation settings are as follows: For the methods that have Tikhonov regularized term λ_t and manifold regularized term λ_m , we use 5-fold cross validation to determine the value of λ_t and λ_t . The candidate set for λ_t and λ_t are $\left\{10^{-6},10^{-3},10^{-2},10^{-1},1,10,10^2,10^3,10^6\right\}$; For LLGDI and FME, there is an addition parameter, i.e. λ_r that balancing tradeoff between the regression form of Eq. (24) and other terms, we also determine the value using 5-fold cross validation with the same candidate set; for the number of neighborhoods, we choose the same values as in other methods in each dataset. The training set in all datasets are preliminarily processed with PCA operator to eliminate the null space before

performing dimensionality reduction. For supervised methods such as RLDA and RLS, we use only labeled set to train the learner. For semi-supervised dimensionality reduction methods, we use all the training set with both labeled and unlabeled set to train the learner. All algorithms used labeled set in the output reduced space to train a nearest neighborhood classifier for evaluating the classification accuracy of test set.



(a) (b) (c)

Fig. 1 Gray image of reduced space learned by LDA, SDA and LLGDI: two-cycle dataset (a) LDA (b) SDA (c) LLGDI

Table 2

Dataset Information and Data Partition for Each Dataset ('**Balance**' is defined as the ratio between the number of samples in the smallest class and the number of samples in the latest class)

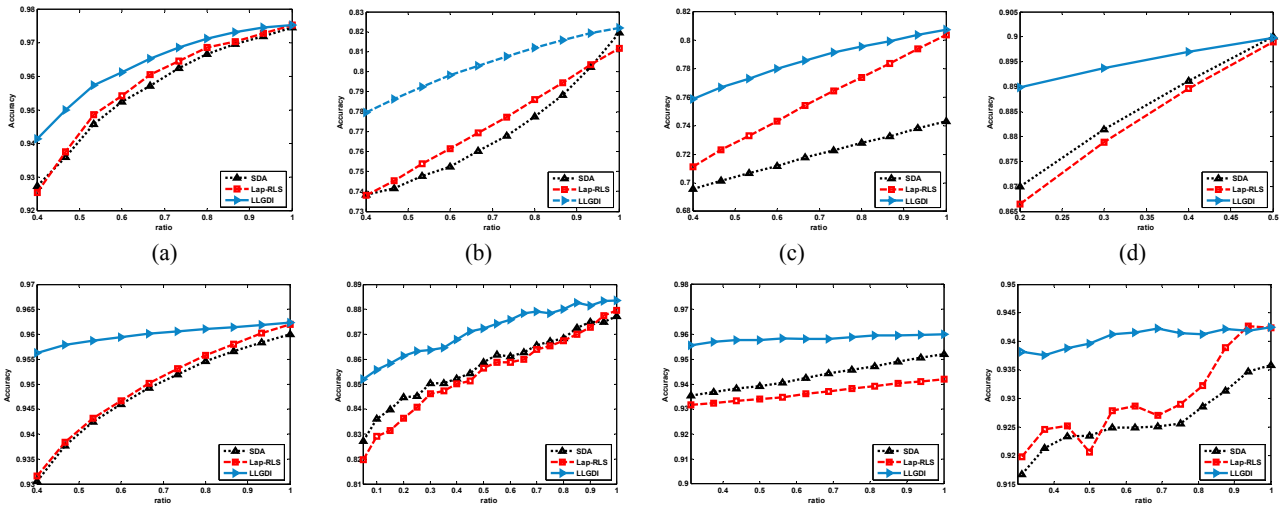
Dataset	Database Type	#Samples(N)	#Dim(D)	#Class(c)	**Balanc**	#Labeled(l	#Unlabeled(u	#Test(t)
UMNIST	Face	1012	1024	20	0.2857	3	8	Remains
CMU-PIE	Face	11554	1024	68	0.9647	5	10	Remains
YALE-B	Face	2414	1024	38	0.9219	5	10	Remains
MIT	Face	3240	1024	10	1	2	8	30
COIL100	Object	7200	1024	100	1	10	20	30
ETH80	Object	3280	1024	80	1	10	20	Remains
USPS	Hand-written-digi	9298	256	10	0.4559	20	80	100
MNIST	Hand-written-digi	60000	784	10	1	20	80	100

Table 3

Average Classification Accuracy over 20 Random Splits on Test Set of Different Datasets (Values in the brackets show the standard derivations)

Dataset	1NN	RLS	RLDA	SDA1	SDA2	LapRLS	LS-SDA	FME	LLGDI
UMNIST	87.99(2.77)	91.50(2.56)	91.51(2.56)	92.74(3.59)	92.55(1.17)	92.55(1.17)	92.55(1.17)	93.65(1.17)	**94.15(1.38)**
CMU-PIE	55.34(1.58)	71.39(1.60)	71.39(1.60)	73.83(2.20)	73.80(2.22)	73.80(2.22)	73.80(2.22)	75.80(6.56)	**77.95(4.52)**
YALE-B	52.65(0.60)	65.70(3.85)	64.66(2.97)	69.57(1.21)	71.15(1.64)	71.15(1.64)	71.15(1.64)	73.85(1.43)	**75.89(1.64)**
MIT	81.17(2.87)	83.71(1.86)	84.64(1.92)	87.00(4.66)	86.65(0.42)	86.65(0.42)	86.65(0.42)	88.92(3.69)	**88.98(3.32)**
COIL100	89.89(1.33)	90.78(2.18)	90.58(1.23)	93.05(4.69)	93.16(1.47)	93.16(1.47)	93.16(1.47)	94.99(3.61)	**95.63(2.62)**
ETH80	67.96(3.16)	71.90(5.43)	72.88(4.22)	82.73(3.05)	81.98(4.20)	81.98(4.20)	81.98(4.20)	84.35(3.04)	**85.19(2.81)**
USPS	90.84(5.82)	91.20(5.60)	91.28(5.25)	93.55(5.12)	93.16(6.25)	93.16(6.25)	93.16(6.25)	94.10(2.47)	**95.56(5.47)**
MNIST	89.34(8.48)	90.58(1.23)	90.18(2.18)	91.67(1.18)	91.98(0.99)	91.98(0.99)	91.98(0.99)	92.98(0.99)	**93.76(1.36)**

The average accuracies over 20 random splits with the above parameters for each dataset are shown in Table 3. From the simulation results, we can obtain the following observation: 1) the semi-supervised dimensionality reduction methods are better than the corresponding supervised methods. For example, SDA1, Lap-RLS outperform RLDA, RLS by about 5%-10% in the Yale-B and ETH80 datasets. For other datasets, it can outperform by 2%-3%. This indicates that by incorporating the unlabeled set into the training procedure, the classification performance can be markedly improved, as the manifold structure embedded in the dataset is preserved. 2) LLGDI and FME deliver the accuracies much better than those delivered by other semi-supervised dimensionality reduction methods such as SDA, Lap-LDA, Lap-RLS by about 2%-3% in most datasets. The improvement can even achieve almost 5% in CMU-PIE dataset for LLGDI. This enhancement is believed to be true due to the reason that by taking into account the local discriminative information, LLGDI preserves more discriminative information which is good for classification. In addition, we find that LLGDI can deliver better result to FME in most datasets. 3) SDA2, Lap-RLS and LS-SDA can achieve the same accuracies due to reason as analyzed in Section III. 4) We also evaluate LLGDI and compare it with SDA1 and Lap-RLS by fixing the number of training set and increasing the number of labeled set. The simulation results can be seen in Fig. 3. Following Fig. 3, we can observe that with the increase of labeled samples, the accuracies of three methods are all improved. However, LLGDI is more robust to the increase of labeled samples, specifically in COIL100, USPS and MNIST datasets. Another observation is that LLGDI can achieve better performances than SDA1 and Lap-RLS given few labeled samples. The reason for it is LLGDI incorporates local discriminative information into learning hence is more robust to the number of labeled samples.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 2 Accuracy v.s. the ratio between the numbers of labeled set and that of training set: (a) UMNIST dataset (b) CMU-PIE dataset (c) Yale-B dataset (d) MIT dataset (e) COIL100 dataset (f) ETH80 dataset (g) USPS dataset (h) MNIST dataset

VI. Conclusion

Semi-supervised dimensionality reduction is an important research topic in many pattern recognition and machine learning applications. Among all the methods for semi-supervised dimensionality reduction, SDA and LapRLS are two popular ones. Though the two methods are actually the extensions of different supervised methods, we show in this paper that they can be unified into a regularized least square framework. However, the regularization term added to the framework focuses on smoothing only, it cannot fully utilize the underlying discriminative information which is vital for classification. In this paper, we propose a new effective semi-supervised dimensionality reduction method, called LLGDI, to solve the above problem. The proposed LLGDI method introduces a discriminative manifold regularization term by using the local discriminative information instead of only relying on neighborhood information. In this way, both the local geometrical and discriminative information of dataset can be preserved. Theoretical analysis and extensive simulations show the effectiveness of the method compared with other existing methods.

Appendix

A. Proof of Theorem 2

Let the rank of auxiliary matrix M be q, then, $U \in R^{c\times q}$ is an orthogonal matrix formed by the eigenvectors corresponding to the q nonzero eigenvalues of M. Since $q \le c$, there exists an orthogonal matrix $U_\perp \in R^{c\times(c-q)}$ formed by the eigenvectors corresponding to the $c-q$ zero eigenvalues of M, which satisfies $U^T U_\perp = O$ and follows:

$$U_\perp^T M U_\perp = U_\perp^T H_b{}^T \left(S_t + \lambda_m XLX^T + \lambda_t I\right)^{-1} H_b U_\perp = O \to H_b U_\perp = O . \tag{27}$$

The third equation holds as $\left(S_t + \lambda_m XLX^T + \lambda_t I\right)$ is a positive definite matrix. Recalling the optimal solution of LS-LDA is $V_{LS-SDA}^* = \left(S_t + \lambda_m XLX^T + \lambda_t I\right)^{-1} H_b$, following Eq. (27), we have $V_{LS-SDA}^* U_\perp = \left(S_t + \lambda_m XLX^T + \lambda_t I\right)^{-1} H_b U_\perp = O$. Hence, if we let $\widetilde{U} = \left[U, U^\perp\right] \in R^{c\times c}$ be the orthogonal matrix satisfying $\widetilde{U}^T \widetilde{U} = \widetilde{U}\widetilde{U}^T = I$, we then have

$$\begin{aligned} V_{LS-SDA}^* V_{LS-SDA}^{*T} &= V_{LS-SDA}^* \widetilde{U}\widetilde{U}^T V_{LS-SDA}^{*T} \\ &= V_{LS-SDA}^* U U^T V_{LS-SDA}^{*T} + V_{LS-SDA}^* U_\perp U_\perp^T V_{LS-SDA}^{*T} \\ &= V_{LS-SDA}^* U U^T V_{LS-SDA}^{*T} \end{aligned} \tag{28}$$

Since $V_{SDA}^* = V_{LS-SDA}^* U$, following Eq. (28), we can prove $V_{LS-SDA}^* V_{LS-SDA}^{*T} = V_{SDA}^* V_{SDA}^{*T}$.

B. Proof of Theorem 4

To prove Theorem 5, we first give two lemmas:

Lemma 1: *Given $A \in R^{n\times n}$ and $B \in R^{n\times n}$ are two positive semi-definite matrixes and the rank of B is t, the for $P \in R^{n\times t}$, we have:*

$$\max_P Tr\left(\left(P^T AP\right)^{-1} P^T BP\right) = \max Tr\left(A^{-1}B\right).$$

The proof of Lemma 1 can be found in [14].

Lemma 2: *For any matrix A, $A\left(A^T A + \lambda I\right)^{-1} = \left(AA^T + \lambda I\right)^{-1} A$ holds.*

Proof of Lemma: Note $A\left(A^T A + \lambda I\right) = \left(AA^T + \lambda I\right) A$. Then,

$$\begin{aligned} A\left(A^T A + \lambda I\right)^{-1} &= \left(AA^T + \lambda I\right)^{-1}\left(AA^T + \lambda I\right) A\left(A^T A + \lambda I\right)^{-1} \\ &= \left(AA^T + \lambda I\right)^{-1}\left(AA^T + \lambda I\right) A\left(A^T A + \lambda I\right)^{-1} \\ &= \left(AA^T + \lambda I\right)^{-1} A\left(A^T A + \lambda I\right)\left(A^T A + \lambda I\right)^{-1} \\ &= \left(AA^T + \lambda I\right)^{-1} A \end{aligned}$$

Proof of Theorem 4: According to Lemma 1, we have

$$J(V) = \max Tr\left(\left(XX^T + \lambda I\right)^{-1} XG^T GX^T\right).$$

According to the trace property, i.e. $Tr(AB)=Tr(BA)$ and Lemma 2, we have

$$\begin{aligned} J(V) &= \max Tr\left(GX^T\left(XX^T+\lambda I\right)^{-1}XG^T\right) \\ &= \max Tr\left(GX^TX\left(X^TX+\lambda I\right)^{-1}G^T\right) \\ &= \max Tr\left(G\left(X^TX+\lambda I-\lambda I\right)\left(X^TX+\lambda I\right)^{-1}G^T\right) \\ &= \max Tr\left(GG^T-\lambda G\left(X^TX+\lambda I\right)^{-1}G^T\right) \\ &= \min Tr\left(G\left(X^TX+\lambda I\right)^{-1}G^T\right) \end{aligned}$$

The last equation holds as $GG^T=I$, we thus prove Theorem 4.

C. Derivation of Eq. (23)

Let $A^T=\left(XL_cX^T+\lambda I\right)^{-1}XL_c$, following Eq. (21) and (22), we have $\left\|V^TX+b^Te-G\right\|_F^2+\lambda\|V\|_F^2=\left\|G\left(N-L_c\right)\right\|_F^2+\lambda\|GA\|_F^2$, which can be derivate as

$$\begin{aligned} &\left\|G\left(N-L_c\right)\right\|_F^2+\lambda\|GA\|_F^2 \\ &=Tr\left(GNN^TG^T\right)-2Tr\left(GNG^T\right)+Tr\left(GL_cG^T\right)+\lambda Tr\left(GAA^TG^T\right) \\ &=Tr\left(GNG^T\right)-2Tr\left(GNG^T\right)+Tr\left(GL_cG^T\right) \\ &=Tr\left(G\left(L_c-N\right)G^T\right) \end{aligned} \quad (29)$$

The second equation holds $L_c=L_cL_c$ and the third equation holds as

$$\begin{aligned} &\lambda Tr\left(GAA^TG^T\right)+Tr\left(GNN^TG^T\right) \\ &=Tr\left(GL_cX^T\left(XL_cX^T+\lambda I\right)^{-1}\left(XL_cX^T+\lambda I\right)^{-1}XL_cG^T\right)+ \\ &\quad Tr\left(GL_cX^T\left(XL_cX^T+\lambda I\right)^{-1}XL_cX^T\left(XL_cX^T+\lambda I\right)^{-1}XL_cG^T\right) \\ &=Tr\left(GL_cX^T\left(XL_cX^T+\lambda I\right)^{-1}\left(XL_cX^T+\lambda I\right)\left(XL_cX^T+\lambda I\right)^{-1}XL_cG^T\right) \\ &=Tr\left(GL_cX^T\left(XL_cX^T+\lambda I\right)^{-1}XL_cG^T\right) \\ &=Tr\left(GNG^T\right) \end{aligned}$$

In addition, we have:

$$\begin{aligned} L_c-N &= L_c-L_cL_cX^T\left(XL_cX^T+\lambda I\right)^{-1}XL_cL_c \\ &= L_c-L_cL_cX^TXL_c\left(L_cX^TXL_c+\lambda I\right)^{-1}L_c \\ &= L_c-L_c\left(L_cX^TXL_c+\lambda I-\lambda I\right)\left(L_cX^TXL_c+\lambda I\right)^{-1}L_c \\ &= L_c-L_c\left(I-\left(L_cX^TXL_c+\lambda I\right)^{-1}\right)L_c \\ &= L_c\left(L_cX^TXL_c+\lambda I\right)^{-1}L_c \end{aligned} \quad (31)$$

We then finish the derivation of Eq. (23).

ACKNOWLEDGEMENT

The work was supported in part by the Shenzhen Foundation Research Fund under Grant no. JCY20120613115205826 and the Shenzhen Strategic Emerging Industries Program under Grant no. ZDSY20120613125016389.

REFERENCES

[1] T. Zhang, D. Tao X. Li and J. Yang, "Patch alignment for dimensionality reduction," IEEE Trans. on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1299-1313, 2009

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

[3] M. Turk, A. Pentland, "Face recognition using Eigenfaces," In Proc. of CVPR, 1991

[4] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, Eigenfaces vs. Fisherfaces: "Recognition using class specific linear projection", IEEE Trans. on PAMI, vol. 19 no. 7, pp. 711-720, 1997

[5] J. H. Friedman, "Regularized discriminant analysis," Journal of the American Statistical Association, vol. 84, no. 405, pp. 165-175, 1989

[6] X. Zhu, Z. Ghahramani and J. D. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," In Proc. of ICML, 2003

[7] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Scholkopf, "Learning with local and global consistency," In Proc. of NIPS, 2004

[8] D. Cai, X. He and J. Han, "Semi-supervised discriminant analysis," In Proc. of ICCV, 2007

[9] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: a geometric framework for learning from labeled and unlabeled examples,", Journal of Machine Learning Research, vol. 7, pp. 2399-2434, 2006

[10] X. He, S. Yan, Y. Hu, P. Niyogi and H. Zhang, "Face recognition using Laplacianfaces", IEEE Trans. on PAMI, vol. 27, no. 3, pp. 328-340, 2005

[11] J. Ye, "Least square linear discriminant analysis," In Proc. of ICML, 2007

[12] Z. Zhang, G. Dai, C. Xu and M. I. Jordan, "Regularized Discriminant Analysis, Ridge Regression and Beyond", Journal of Machine Learning Research, vol. 11, pp. 2199-2228, 2010

[13] L. Sun, B. Ceran and J. Ye, "A scalable two-stage approach for a class of dimensionality reduction techniques," In Proc. of KDD, 2010

[14] K. Fukunaga, "Introduction to Statistical Pattern Classification," Academic Press, 1990

[15] C. Zhang, F. Nie and S. Xiang, "A general kernelization framework for learning algorithms based on kernel PCA," Neurocomputing, vol. 73, no. 4-6, pp. 959-967, 2010

[16] F. Nie, D. Xu, I.W.H. Tsang, C. Zhang, "Flexible manifold embedding: a framework for semi-supervised and unsupervised dimension reduction," IEEE Trans. on Image Processing, vol. 19, no. 7, pp. 1921-1932, 2010

[17] Y. Yang, D. Xu, F. Nie, S. Yan and Y. Zhuang, "Image clustering using local discriminant models and global integration," IEEE Trans. on Image Processing, vol. 19, no. 10, pp. 2761-2773, 2010

[18] Zhao Zhang, Tommy W. S. Chow, and Mingbo Zhao, "Trace Ratio Optimization based Semi-Supervised Multimodal Nonlinear Dimensionality Reduction for Marginal Manifold Visualization," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 5, pp. 1148-1161, 2013

[19] Zhao Zhang, Mingbo Zhao, and Tommy W. S. Chow, "Marginal Semi-Supervised Sub-Manifold Projections with Informative Constraints for Dimensionality Reduction and Recognition," Neural Networks, vol. 36, pp. 97-111, 2012

[20] Zhao Zhang, Mingbo Zhao, and Tommy W. S. Chow, "Constrained Large Margin Local Projection Algorithms and Extensions for Multimodal Dimensionality Reduction. Pattern Recognition," vol. 46, no. 12, pp. 4466-4493, 2012

[21] Zhao Zhang, Tommy W. S, Chow, and Mingbo Zhao, "M-Isomap: Orthogonal Constrained Marginal Isomap for Nonlinear Dimensionality Reduction," IEEE Transactions on Systems, Man and Cybernetics Part B: Cybernetics, vol. 43, no.1, pp. 180-192, 2013

[22] Zhao Zhang, Mingbo Zhao, and Tommy W. S. Chow, "Binary- and Multi-Class Group Sparse Canonical Correlation Analysis for Feature Extraction and Classification," accepted in TKDE, 2012

[23] Mingbo Zhao, Zhao Zhang, Tommy W. S. Chow, "Trace ratio criterion based generalized discriminative learning for semi-supervised dimensionality reduction," Pattern Recognition, vol. 45, no. 4, pp. 1482-1499, 2012

[24] Mingbo Zhao, Zhao Zhang, Haijun Zhang, "Soft Label based Linear Discriminant Analysis for Semi-supervised Dimensionality Reduction," In Proc. of IJCNN, 2013