Pattern Recognition 67 (2017) 252–262

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Low rank representation with adaptive distance penalty for semi-supervised subspace classification

Lunke Fei[a], Yong Xu[b,*], Xiaozhao Fang[b], Jian Yang[c]

[a] School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China
[b] Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China
[c] College of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing, China

ARTICLE INFO

Article history:
Received 13 January 2016
Revised 3 November 2016
Accepted 11 February 2017
Available online 13 February 2017

Kewwords:
Low rank representation
Adaptive distance penalty
Similarity graph construction
Semi-supervised classification
Projection of LRRADP

ABSTRACT

The graph based Semi-supervised Subspace Learning (SSL) methods treat both labeled and unlabeled data as nodes in a graph, and then instantiate edges among these nodes by weighting the affinity between the corresponding pairs of samples. Constructing a good graph to discover the intrinsic structures of the data is critical for these SSL tasks such as subspace clustering and classification. The Low Rank Representation (LRR) is one of powerful subspace clustering methods, based on which a weighted affinity graph can be constructed. Generally, adjacent samples usually belong to a union of subspace and thereby nearby points in the graph should have large edge weights. Motivated by this, in this paper, we proposed a novel LRR with Adaptive Distance Penalty (LRRADP) to construct a good affinity graph. The graph identified by the LRRADP can not only capture the global subspace structure of the whole data but also effectively preserve the neighbor relationship among samples. Furthermore, by projecting the data set into an appropriate subspace, the LRRADP can be further improved to construct a more discriminative affinity graph. Extensive experiments on different types of baseline datasets are carried out to demonstrate the effectiveness of the proposed methods. The improved method, named as LRRADP^2, shows impressive performance on real world handwritten and noisy data. The MATLAB codes of the proposed methods will be available at http://www.yongxu.org/lunwen.html.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

In many big data related applications, the problem of effectively connecting unlabeled data with labeled data is of central importance [1,2]. For example, in the applications of image based web searching and image based object recognition, the labeled data is usually limited and the unlabeled data are rich and available in internet. In these problems, the target goal is to build the connection between unlabeled data and labeled data and then identify the labels of the unlabeled data. Semi-supervised Subspace Learning (SSL) [3–6] is a family of techniques that exploits the "manifold structure" of the data by using both labeled and unlabeled samples [7,8].

Constructing a graph of the local connectivity of data is an effective strategy for SSL due to its success in practice [9–11]. The graph based SSL methods treat both labeled and unlabeled samples from the data set as nodes in a graph, and then instantiate edges among these nodes which are weighted according to the affinity between the corresponding pairs of samples. Suppose that the data set is noiseless and embeds in independent subspaces, the graph identified by the respective SSL method should be a block-diagonal matrix and each block corresponding to a subspace. Subspace clustering is able to produce exactly correct clustering result based on the block-diagonal matrix. To address this issue, we build a weight graph $G = (V, W)$, where V is the vertex set denoting nodes of the graph corresponding to N data points and $W \in R^{N \times N}$ is a symmetric non-negative weight matrix representing the relationship among the nodes. A non-zero weight reflects the affinity between corresponding nodes and a zero weight denotes that there is no edge jointing them. An ideal similarity matrix W, hence an ideal weight graph G, is one in which nodes that correspond to points from the same subspace are connected to each other and there is no edge between any two nodes that correspond to points belonging to different subspaces. Thus, given a data set, the problem of graph construction is to determine the weight matrix W. A perfect similarity graph built by SSL has n independent connected components corresponding to n subspaces and then by applying spectral clustering the labels can be propagated from the labeled samples to unlabeled samples over the graph [12–15].

* Corresponding author.
E-mail address: yongxu@ymail.com (Y. Xu).

http://dx.doi.org/10.1016/j.patcog.2017.02.017
0031-3203/© 2017 Elsevier Ltd. All rights reserved.

The recently Low Rank Representation (LRR) [13,14] is a promising weight graph construction method. The target of the LRR aims at finding the lowest-rankness representation among all candidates that can express the data vectors as linear combinations of the basis in a proper dictionary. Consider a set of data $X = [x_1, x_2, \ldots, x_n]$ in R^d, each column of which is a sample that can be represented by the linear combination of a basis of d vectors. If we form the basis matrix as $A = [a_1, a_2, \ldots, a_m]$, the X can be represented as:

$$X = AZ, \tag{1}$$

where $Z = [z_1, z_2, \ldots, z_n]$ is the coefficient matrix with each z_i characterizing how other samples contribute to the representation of x_i. Since the data can only be essentially represented by those data in the same subspace, the nonzero elements in z_i represents that the corresponding samples are in the same subspace. Therefore, minimizing rankness of the data vector space could be an appropriate criterion to cluster data drawn from multiple linear subspaces. That is, LRR discovers the lowest rankness of the representation of the data set as follows.

$$\min_Z \ rank(Z), \ s.t. \ X = AZ, \tag{2}$$

where $rank(\bullet)$ denotes the rankness of a matrix. The low rankness constraint guarantees that the coefficients of samples coming from the same subspace are highly correlated. When the data are clean and exactly from linearly independent subspaces, the similarity matrix built by this way is an ideally n block diagonal matrix corresponding to n subspaces.

LRR is an effective framework for exploring the multiple subspace structures of data. Based on the LRR, lots of recent efforts have been made to exploit ways of constructing a discriminative graph for SSL [16–19]. Liu et al. [18] proposed a latent low rank representation method for subspace clustering by approximating and using the unobserved data hidden in the observed data to resolve the issue of insufficient sampling. Zhang et al. [19] extended the latent LRR by choosing the sparest solution in the solution set to increase the robustness of the method. Wei et al. [20] proposed a robust shape interaction by preprocessing the data using robust PCA [21] and then applying LRR to build the similarity matrix. By combining the sparsity and global structure, Zhuang et al. [22,23] proposed a nonnegative low-rank and sparse graph for semi-supervised learning. Fang et al. [24,25] combined the nonnegative low-rank representation with the semi-supervised clustering learning within one framework achieving acceptable classification performance.

Conventional LRR based methods usually consider much on construction of the global subspace structure. However, a good graph should not only capture global structures of all the data but also reveal the intrinsic neighbor relationship among the data [26]. In this paper, we propose a Low Rank Representation with Adaptive Distance Penalty (LRRADP) method, which constructs the linear combination by using the nearby samples as much as possible via the adaptive distance penalty. The affinity graph built by the LRRADP can better both capture the global subspace structure of a whole data set and preserve local neighbor relationships among the data samples. The similarity graph/matrix identified by the LRRADP can work well with conventional semi-supervised classification method, such as Gaussian Fields and Harmonic Functions (GFHF) [8], for the label prediction of unlabeled samples. Moreover, the LRRADP is improved to LRRADP^2 by projecting the data set into an appropriate subspace.

The remainder of this paper is organized as follows. Section 2 introduces the related works of the low rank representation and semi-supervised subspace classification methods. Section 3 proposes an LRR with adaptive distance penalty method (LRRADP) for subspace classification. Section 4 extends the LRRADP to LRRADP^2 by projecting the data into an appropriate subspace. Section 5 presents the experimental results and Section 6 concludes this paper.

2. Related works

2.1. Low rank representation

To capture the global structure of data, LRR [13,14] is to construct the affinities of an undirected graph. A LRR graph obtains the representation of all data under a global low-rank constraint, thus is better at capturing the global data structures. It has been proven that under suitable conditions, LRR can correctly preserve the membership of samples that belong to the same subspace [13]. Given a set of data, the data usually can be represented by other data that lie in the same subspace. When the subspace are independent and the data is noiseless. The subspace can be exactly divided and the representation of the data set presents block diagonal. The LRR demonstrated that minimizing rank representation of the data set can be replaced by minimizing the nuclear norm of the union data, resulting in the following low rank optimization problem:

$$\min_Z \ ||Z||_*, \ s.t. \ X = AZ, \tag{3}$$

where $\|\bullet\|_*$ denotes the nuclear norm of a matrix which equals to the sum of the singular values of the matrix. A is a basis matrix that is used as the dictionary of the linear representation. By choosing an appropriate dictionary A, the underlying row space of the data set X can be correctly captured. In most conditions, the data matrix itself X is directly used as the dictionary [13]. We call the optimal solution of the problem (3) as the "lowest rank representation" of data X. In real-world applications, the observation data are often corrupted by noise. By correcting the noise, the model of LRR can be converted to as follows.

$$\min_Z \ ||Z||_* + \lambda||E||_l, \ s.t. \ X = XZ + E, \tag{4}$$

where $\lambda > 0$ is a parameter. E is an error matrix representing the noises and $\|\bullet\|_l$ denotes a special regularization strategy to characterize the noise. There are many choices to define the error term. For example, $||E||_F^2$ is proposed for the small Gaussian noise, $\|E\|_0$ can be used to character random corruptions, and $\|E\|_{2,1}$ generally character "sample-specific corruption" by encouraging the columns in E to be zero.

Based on the basic low rank representation, a lot of efforts have been devoted to improve the LRR by imposing the penalty on Z and E and different kinds of LRR based methods were proposed. To better handle the LRR based method, a more general rank minimization problem is given as follows.

$$\min_Z \ ||Z||_* + Q(Z, E), \ s.t. \ X = XZ + E, \tag{5}$$

where Q is a penalty function on Z and E. For example, Zhuang et al. [22] defined the $Q(Z, E) = \lambda_1||Z||_1 + \lambda_2||E||_{2,1}$, where λ_1 and λ_2 are non-negative parameters, to construct a low-rank and sparse graph. Feng et al. [27] aim at producing a exactly block-diagonal similarity matrix by restricting the rank of Laplacian matrix by defining $Q(Z, E) = \frac{\lambda}{2}||E||_F^2$.

2.2. Semi-supervised classification

Based on the affinity graph/matrix obtained by respective subspace clustering methods, semi-supervised classifier, such as Local and Global Consistency (LGC) [28] and GFHF [8], can be used to predict labels of unlabeled samples. To address this issue, we define a matrix $F = [F_l F_u]^T \in R^{n \times c}$ to represent the label prediction matrix by labeling a sample x_i with a label $y_i = \arg\max_j F_{i,j}$. Let

$Y = [Y_l Y_u]^T \in R^{n\times c}$ be a labeled matrix, where Y_l and Y_u correspond to the labeled and unlabeled samples, respectively. $Y_{i,j} = 1$ if sample x_i is associated with label $j (j = 1, 2, \ldots, c)$ and $Y_{i,j} = 0$ otherwise. Both LGC and GFHF utilize the weight graph and labeled matrix to recover the continuous classification function by optimizing both the label fitness and manifold smoothness. In other words, F should satisfy the given labels Y_l and meanwhile smooth on the whole graph built based on both labeled and unlabeled samples. That is, LGC and GFHF aim at minimizing the following optimization cost on a weight graph to recover the classifiers F, respectively.

$$g_{LGC}(F) = \frac{1}{2}\sum_{i,j=1}^{n} ||\frac{F_{i*}}{\sqrt{D_{ii}}} - \frac{F_{j*}}{\sqrt{D_{jj}}}||_2^2 S_{i,j} + \lambda \sum_{i=1}^{m} ||F_{i,*} - Y_{i,*}||^2,$$
$$g_{GFHF}(F) = \frac{1}{2}\sum_{i,j=1}^{n} ||F_{i*} - F_{j*}||_2^2 S_{i,j} + \lambda_\infty \sum_{i=1}^{m} ||F_{i,*} - Y_{i,*}||^2, \quad (6)$$

where λ balances the label fitness and manifold smoothness. λ_∞ is a very large number so $\sum_{i=1}^{n} ||F_i - Y_i||^2$ must be very small. $F_{i,*}$ and $Y_{i,*}$ are the ith row of F and Y, respectively. m is the number of labeled samples. Both $g_{LGC}(F)$ and $g_{GHF}(F)$ have following similar formulations:

$$g_{LGC}(F) = tr(F^T \tilde{L} F) + tr(F - Y)^T U_\lambda (F - Y),$$
$$g_{GFHF}(F) = tr(F^T L F) + tr(F - Y)^T U_\infty (F - Y), \quad (7)$$

where $\tilde{L}$ and L are normalized Laplacian matrix and Laplacian matrix of the similarity matrix S, respectively. U_λ is a diagonal matrix with the m elements as λ corresponding to labeled samples and with the rest $n - m$ diagonal elements as 0 corresponding to unlabeled samples, respectively. U_∞ is also a diagonal matrix with the m elements as λ_∞ corresponding to labeled samples and with the rest $n - m$ diagonal elements as 0 corresponding to unlabeled samples, respectively. F can be directly solved by differentiating g_{LGC} or g_{GFHF} with respect to F.

3. Low rank representation with adaptive distance penalty

Throughout this paper, all the matrices are written as uppercase. For matrix M, the (i, j)th element of M is denoted as $[M]_{i,j}$. The ith row of M is denoted as $[M]_{i,*}$ and the jth column of M is denoted as $[M]_{*,j}$. The trace of M is denoted as $tr(M)$. The $l_p - norm$ of M is denoted as $||M||_p$. Specially, the Frobenius norm and nuclear norm of matrix M are denoted as $||M||_F$ and $||M||_*$, respectively. The transpose of M is denoted as M^T. $M \geq 0$ mean all elements of M are larger than or equal to zero. I denotes an identity matrix. In this section, we use $X \in R^{d\times n}$ to represent the data set, where d is the dimension of the data and n is the number of the data.

3.1. LRRADP

The graph identified by LRR obtains the representation of all the data under a global low-rank constraint, and thus is better at capturing the global structures of data, such as multiple clusters and subspaces.

Intuitively, nearby points are possibly from the same subspace and thus the unlabeled nearby points in graph should have similar labels. Motivated by this, the following quadratic energy function is suitable to determine the weight between corresponding pairs of points in the graph:

$$\sum_{i,j} ||[X]_{*,i} - [X]_{*,j}||_2^2 [Z]_{i,j}, \quad (8)$$

where $[Z]_{i,j}$ is the related weight between the ith and jth samples. Minimizing (8) can assign small weight to the edge between samples with far distance. On the other hand, nearby points in the graph might obtain relatively large edge weight. By combining (8) with the LRR, the proposed Low Rank Representation with Adaptive Distance Penalty (LRRADP) method is defined as follows.

$$\min_Z ||Z||_* + \lambda \sum_{i,j=1}^{n} ||[X]_{*,i} - [X]_{*,j}||_2^2 [Z]_{i,j}, \ s.t\ X = XZ, \quad (9)$$

where Z is the low rank representation matrix of the data set. $\lambda > 0$ is a balance parameter. By setting an appropriate λ, the LRRADP can better capture the global subspace structure of a union space and preserve the neighborhood relativity between nearby points. In (9), a**s** the coefficients can be negative in the data representation of the LRR, which allows the data can be substracted by each other [22]. It lacks physical interpretation for many real applications [21,23]. To address this issue, we impose the nonnegative constraint on the data representation. In addition, we simplify the symbols by replacing the element operation by matrix operation and the LRRADP can be reformulated as

$$\min_{Z,E} ||Z||_* + \lambda tr(\Xi(D \otimes Z)), \ s.t\ X = XZ, \ Z \geq 0. \quad (10)$$

where $D \in R^{n\times n}$ is the distance matrix of the X, in which $[D]_{i,j} = ||[X]_{*,i} - [X]_{*,j}||_2^2$ is the distance of two points. $\Xi \in R^{n\times n}$ is a matrix with all elements are 1. "$\otimes$" is the Hadamand product.

In real-world applications, data are often noisy due to measurement or processing issue. In such cases, the data do not perfectly lie in a union of subspaces. So we need to consider the case where data X is a noisy matrix. We introduce an error matrix E to model the noise, resulting in the following optimization problem:

$$\min_{Z,E} ||Z||_* + \lambda_1 ||E||_p + \lambda_2 tr(\Xi(D \otimes Z)), \ s.t\ X = XZ + E, \ Z \geq 0. \quad (11)$$

The objective function of the LRRADP has three terms. The first term pursues the lowest rank representation of the whole data space. The second term characterizes the noises of the data and third term ensures that nearby points in Euclidean space can be assigned relatively large edge weights. $\lambda_1 > 0$ and $\lambda_2 > 0$ are balance parameters to trade off among the low rankness representation, errors and adaptive distance penalty. In this paper, $l_1 - norm$ of E is used to characterize sparse noise of the data. So the LRRADP has the following model.

$$\min_{Z,E} ||Z||_* + \lambda_1 ||E||_1 + \lambda_2 tr(\Xi(D \otimes Z)), \ s.t\ X = XZ + E, \ Z \geq 0. \quad (12)$$

The minimizer Z^* of the LRRADP can be considered as an improved low rank representation by embedding the adaptive distance penalty, a column of which naturally characterizes the affinity of a sample with other samples. Since the adaptive distance penalty generally guarantees that the representation coefficients of a sample mainly be formed by the neighboring samples, the weight graph identified by the optimal solution of the LRRADP can better capture both the global clustering structure of the whole data and local neighbor relationships among the samples. Note here that, since each sample can be represented by itself, the LRRADP always exist feasible solutions even when the data sampling is insufficient.

3.2. Solving the LRRADP

To solve the LRRADP, we first introduce an auxiliary variable H to make the variables in the LRRADP separable and then reformulate problem (5) as follows.

$$\min_{Z,E} ||Z||_* + \lambda_1 ||E||_1 + \lambda_2 tr(\Xi(D \otimes H)),$$
$$s.t\ X = XZ + E, \ Z = H, H \geq 0. \quad (13)$$

This problem can be solved by using the Augmented Lagrange Multiplier (ALM) method [30–32]. The augmented Lagrange function of problem (13) is

$$\begin{aligned} &L(Z,E,H,Y_1,Y_2,\beta)\\ &= ||Z||_* + \lambda_1||E||_1 + \lambda_2 tr(\Xi(D\otimes H)) + \langle Y_1, X - XZ - E\rangle\\ &\quad + \langle Y_2, Z - H\rangle + \frac{\beta}{2}(||X - XZ - E||_2^2 + ||Z - H||_2^2), \end{aligned} \tag{14}$$

where Y_1 and Y_2 are two Lagrange multipliers, $\langle \bullet, \bullet\rangle$ represents the inner product, and $\beta > 0$ is a penalty parameter. By relaxing the inner product, (14) can be further converted to as follows.

$$\begin{aligned} L(Z,E,H,Y_1,Y_2,\beta) &= ||Z||_* + \lambda_1||E||_1 + \lambda_2 tr(\Xi(D\otimes H))\\ &\quad + \frac{\beta}{2}(||X - XZ - E + Y_1/\beta||_2^2 + ||Z - H + Y_2/\beta||_2^2)\\ &\quad - \frac{1}{2\beta}(||Y_1||_2^2 + ||Y_2||_2^2). \end{aligned} \tag{15}$$

The problem (15) is unconstrained. So it can be minimized with respect to Z, E, H, respectively, by fixing the other variables, and then updating the Lagrange multipliers Y_1 and Y_2. Particularly, the update of Z, E and, H in (8) go as follows.

$$\begin{aligned} Z_{k+1} = \arg\min_Z ||Z||_* + \frac{\beta}{2}\Big(&||X - XZ - E_k + \frac{Y_{1,k}}{\beta}||_2^2 + ||Z\\ &- H_k + \frac{Y_{2,k}}{\beta}||_2^2\Big), \end{aligned} \tag{16}$$

$$E_{k+1} = \arg\min_E ||E||_1 + \frac{\beta}{2}||X - XZ_k - E + Y_{1,k}/\beta||_2^2, \tag{17}$$

$$H_{k+1} = \arg\min_H \lambda_2 tr(\Xi(D\otimes H)) + \frac{\beta}{2}||Z_k - H + Y_{2,k}/\beta||_2^2. \tag{18}$$

For problem (16), suppose that $q(Z) = \frac{\beta}{2}(||X - XZ - E_k + \frac{Y_{1,k}}{\beta}||_2^2 + ||Z - H_k + \frac{Y_{2,k}}{\beta}||_2^2)$. By linearizing the quadratic term in (9) at X_k and adding a proximal term, it can be led to the following approximation.

$$\begin{aligned} Z_{k+1} &= \arg\min_Z ||Z||_* + q(Z_k) + \langle \nabla_Z q, Z - Z_k\rangle + \frac{\beta\eta_z}{2}(||Z - Z_k||_2^2)\\ &= \arg\min_Z ||Z||_* + \frac{\beta\eta_z}{2}||Z - Z_k\\ &\quad + \frac{1}{\eta_z}(-X^T(X - XZ - E + Y_1/\beta) + (Z - H + Y_2/\beta))||_2^2, \end{aligned} \tag{19}$$

where $\nabla_Z q$ is the partial differential of q with respect to Z. The solution of the (12) can be obtained by using the Singular Value Thresholding (SVT) operator [33]:

$$\begin{aligned} Z_{k+1} = \Phi_{\frac{1}{\eta_z\beta}}\Big(&Z_k + \frac{1}{\eta_z}(X^T(X - XZ_k - E_k + Y_{1\,k}/\beta)\\ &- (Z_k - H_k + Y_{2\,k}/\beta))\Big), \end{aligned} \tag{20}$$

where Φ represents the SVT operator and $\frac{1}{\eta_z\beta}$ is the thresholding value of the Φ.

For problem (17), it can be directly solved via the shrinkage operator [34].

$$E_{k+1} = \Psi_{\frac{\lambda_1}{\beta}}(X - XZ + Y_1/\beta), \tag{21}$$

where Ψ represents the shrinkage operator and $\frac{\lambda_1}{\beta}$ is the shrinkage threshold of the Ψ.

For problem (18), note that the solutions of different samples are independent, the solution of (18) can be calculated by decomposing it into n independent sub-problems, each of which has a closed form solution.

Algorithm 1

Input: the data set X, parameters: $\lambda_1 > 0$, $\lambda_2 > 0$
Internalize: $Z = H = Y_2 = 0$, $E = Y_1 = 0$, $\beta_0 = 1$, $\beta_{\max} = 10^4$, $\eta_Z = 2||X||^2$, $\xi = 10^{-5}$, $\rho = 1.01$, $k = 0$.
while $||Z_{k+1} - Z_k||/||Z_k|| \geq \xi$
  Update Z as (12);
  Update E as (13);
  Update H as (14);
  Update Y_1 as $Y_{1\,k+1} = Y_{1\,k} + \beta_k(X_{k+1} - X_{k+1}Z_{k+1} - E_{k+1})$;
  Update Y_2 as $Y_{2\,k+1} = Y_{2\,k} + \beta_k(Z_{k+1} - H_{k+1})$;
  Update β as $\beta_{k+1} = \min(\beta_{\max}, \rho\beta_k)$;
  Update k: $k = k + 1$;
end while
Output: an optimal solution (Z^*, H^*, E^*)

$$\begin{aligned} &[H_{k+1}]_{*,i}\\ &= \arg\min_{[H]_{*,i}} \lambda_2[D]_{*,i}^T[H]_{*,i} + \frac{\beta}{2}||[Z_k]_{*,i} - [H]_{i,*} + [Y_{2\,k}]_{*,i}/\beta||_2^2\\ &= \arg\min_{[H]_{*,i}} ||[Z_k]_{*,i} - [H]_{*,i} + [Y_{2\,k}]_{*,i}/\beta - \lambda_2[D]_{*,i}/\beta||_2^2\\ &= [Z_k]_{*,i} + [Y_{2\,k}]_{*,i}/\beta - \lambda_2[D]_{*,i}/\beta \quad (i = 1,\ldots,n). \end{aligned} \tag{22}$$

The complete algorithm of the LRRADP is outlined in Algorithm 1. The algorithm of the LRRADP shares similar convergence propertie as the LADMAP method [29]. Since the η_Z is initialized as larger than $||X||^2$, the LRRADP will convergence to an exact solution [29].

The most computational demand of Algorithm 1 is at step 1, which computes the SVD of matrices. It is easy to check that the computation cost of step 1 is $O(n^3 + d^3)$, where n and d are the sample number and dimension of the dataset. Thus, the above algorithm can be solved with a computation complexity of $O(\zeta(n^3 + d^3))$, where ζ the maximum iteration number of the Algorithm 1.

After obtaining optimal representation matrix Z^*, the weight matrix/graph of the data set can be built as $W = (Z^* + Z^{*T})/2$, and then the semi-supervised classification method, such as LGC and GFHF, can be employed on the weight matrix to conduct classification. In our method, GFHF is used as the label prediction method which has the following optimization form:

$$\begin{aligned} g(F) &= \frac{1}{2}\sum_{i,j=n}^{n} |||[F]_{i,*} - [F]_{j,*}||^2[W]_{i,j} + \lambda_\infty \sum_{i=1}^{m} |||[F]_{i,*,} - [Y]_{i,*,}||^2\\ &= tr(F^T L_W F) + tr(F - Y)^T U(F - Y), \end{aligned} \tag{23}$$

where $F \in R^{n\times c}$ and $Y \in R^{n\times c}$ are the label prediction matrix and labeled matrix, respectively. $L_W = D - W$ is the Laplacian matrix of W, in which D is a diagonal matrix with $D_{ii} = \sum_{j=1}^{n}[W]_{i,j}$. λ_∞ is a large enough parameter. U is a diagonal matrix with diagonal elements are λ_∞ and 0 corresponding to the labeled and unlabeled elements, respectively. By setting the derivative of g with respect to F to zero, the label prediction matrix can be directly obtained as follows.

$$F = (L_W + U)^{-1}UY. \tag{24}$$

Finally, the label of each unknown sample can then be identified as:

$$Label(k) = \arg\max_j F_{k,j}. \tag{25}$$

4. LRRADP²

In general, the quality of data representation will greatly affect the quality of graph. A good data representation could improve the quality of the graph and then improve the performance of the SSL.

Previous works [22] have shown that by projecting the data with a projection matrix, the embedded data will facilitate the subsequent data representation and increase the classification accuracy. To improve the data representation, we propose to learn an appropriate subspace in which the graph identified by the LRRADP is more robust to the variance of data. The adaptive distances between corresponding samples are also calculated in the subspace. We first denote an projection matrix as P. By plugging the learning of P into the LRRADP graph construction framework, we arrive at the following formulation:

$$\min_{Z,P,E} ||Z||_* + \lambda_1||E||_1 + \lambda_2 \sum_{i,j=1}^{n} ||[P^TX]_{*,i} - [P^TX]_{*,j}||_2^2 [Z]_{i,j},$$
$$s.t\ P^TX = P^TXZ + E,\ Z \geq 0, P^TP = I. \tag{26}$$

To avoid the trivial solution, the project matrix P is restricted to column orthogonal. For simplification, we term the problem (26) as LRRADP Projected version (referred to as LRRADP^2). It is easy to check that the LRRADP is a special case of LRRADP^2 with $P = I$. By defining an appropriate linear subspace structure of the data, it is expected to explore a more powerful discriminative graph to perform the semi-supervised classification. Moreover, by suitably assigning the dimension of P, some noise, such as outliers and random corruptions, could be filtered out. Thus, it is believed that the graph identified by the LRRADP^2 should be more discriminative than that identified by the LRRADP method.

To solve the problem (26), we also introduce an auxiliary variable and relax the constraints in (26) to obtain the following augmented Lagrange function:

$$\begin{aligned} &||Z||_* + \lambda_1||E||_1 + \lambda_2 tr(\Xi(D \otimes H)) \\ &+ \langle Y_1, P^TX - P^TXZ - E\rangle + \langle Y_2, Z - H\rangle \\ &+ \langle Y_3, P^TP - I\rangle + \frac{\beta}{2}(||P^TX - P^TXZ - E||_2^2 + ||Z - H||_2^2 \\ &+ ||P^TP - I||_2^2), \end{aligned} \tag{27}$$

where Y_1, Y_2 and Y_3 are Lagrange multipliers, and $\beta > 0$ is a penalty parameter. Following the commonly used strategy in ADM, we alternatively update the unknown variables. Specifically, we first optimize the objective function of the LRRADP^2 with respect to P by fixing Z, E, and H, then we update Z, E, and H while fixing P.

When Z, E and H are fixed, (27) degrade to:

$$\min_P \lambda_2 tr(P^TXL_HX^TP) + tr(Y_1^T(P^TX - P^TXZ - E)) + tr(Y_3^T(P^TP - I)), \tag{28}$$

where $L_H = D_H - (H^T + H)/2$ is the Laplacian matrix of the H, the D_H is defined as the diagonal matrix where ith diagonal element $[D_H]_{i,i} = \sum_j ([H]_{i,j} + [H]_{j,i})/2, (i = 1, 2, \ldots, n)$. The projection matrix P can be directly solved by setting the partial differential of (28) with respect to P to zero.

When P is fixed, the problem of (26) can be converted to

$$\min_{Z,E} ||Z||_* + \lambda_1||E||_1 + \lambda_2 \sum_{i,j=1}^{n} ||[X']_{*,i} - [X']_{*,j}||_2^2 [H]_{i,j},$$
$$s.t\ X' = X'Z + E,\ Z = H, H \geq 0, \tag{29}$$

where $X' = P^TX$. It can be seen that the problem of (29) has a similar objective function as the LRRADP. Thus, the Z, E and H can be solved iteratively by fixing other variables, arriving at following iterations:

$$\begin{aligned} Z_{k+1} = \Phi_{\frac{1}{\eta_Z\beta}}\Big(Z_k + \frac{1}{\eta_z}((P^TX)^T(P^TX - P^TXZ_k - E_k + Y_{1\,k}/\beta) \\ - (Z_k - H_k + Y_{2\,k}/\beta))\Big), \end{aligned} \tag{30}$$

Algorithm 2

Input: the data set X, parameters: $\lambda_1 > 0$, $\lambda_2 > 0$
Internalize: $Z = H = E = Y_1 = Y_2 = 0$, P_0 is initialized by PCA, $\beta_0 = 1$, $\beta_{\max} = 10^4$, $\eta_Z = 2||X||^2$, $\xi = 10^{-5}$, $\rho = 1.01$, $k = 0$.
while $||Z_{k+1} - Z_k||/||Z_k|| \geq \xi$
  Update P as:
    if k==0 $P = P_0$;
    else update P according to (28);
  Update Z as (30);
  Update E as (31);
  Update H as (32);
  Update Y_1 as $Y_{1\,k+1} = Y_{1\,k} + \beta_k(X_{k+1} - X_{k+1}Z_{k+1} - E_{k+1})$;
  Update Y_2 as $Y_{2\,k+1} = Y_{2\,k} + \beta_k(Z_{k+1} - H_{k+1})$;
  Update β as $\beta_{k+1} = \min(\beta_{\max}, \rho\beta_k)$;
  Update k: $k = k + 1$;
end while
Output: an optimal solution (Z^*, H^*, E^*, P^*)

Fig. 1. Some typical examples of different datasets. The first to fifth rows show the some typical images of the COIL20, Extended YaleB, AR, MNIST and C-Cube datasets, respectively.

$$E_{k+1} = \Psi_{\frac{\lambda_1}{\beta}}(P^TX - P^TXZ + Y_1/\beta), \tag{31}$$

$$[H_{k+1}]_{*,i} = [Z_k]_{*,i} + [Y_{2\,k}]_{*,i}/\beta - \lambda_2[D']_{*,i}/\beta, \quad (i = 1, \ldots, n), \tag{32}$$

where $[D']_{i,j} = ||[P^TX]_{*,i} - [P^TX]_{*,j}||_2^2$. We alternatively solve problem (21), (23), (24) and (25) until convergence. The complete process of the optimization of LRRADP^2 is summarized in Algorithm 2. After getting the optimal solution Z^*, we use the similar scheme as that of LRRADP to construct a weight graph and then predict the labels of unlabeled samples by using the GFHF.

5. Experiments

In this section, we evaluate the performance of the proposed methods on baseline databases, as well as other state-of-the-art graph construction methods. We combine the graphs identified by the LRRADP, LRRADP^2 and conventional popular graph construction methods with the GFHF method to perform the semi-supervised classification, and quantitatively evaluate their performance. We test and compare these solvers on six representative data sets, including the COIL20, AR, Extended Yale B, Isolet5, MNIST and C-Cube datasets. Among them, the COIL20 is an object dataset. The AR and Extended Yale B datasets are two face datasets and the Isolet5 is a voice dataset. The MNIST and C-Cube are two real word handwritten datasets of digits and characters, respectively. In the test dataset, Fig 1 shows some typical images selected from these datasets. In order to test the robustness of the proposed methods, we form several corrupted/noisy datasets by adding block corruptions and random noises of different levels on the COIL20 and Extend Yale B datasets. On each dataset, four representative affinity graphs are constructed as baselines by using the LRR [13], LatLRR

Table 1
The accuracy (%) of classification obtained using different methods on the COIL20 dataset.

#Tr	LRR	LatLRR	SSC	Robust LatLRR	LRRADP	LRRADP^2
1	50.68 ± 5.36	43.54 ± 2.29	59.61 ± 6.85	28.40 ± 7.91	**74.93 ± 1.79**	72.42 ± 3.93
2	67.29 ± 3.30	61.81 ± 2.76	68.26 ± 9.15	37.24 ± 2.35	**85.35 ± 0.93**	82.99 ± 3.56
3	73.16 ± 2.86	71.94 ± 2.88	78.70 ± 4.86	70.12 ± 2.73	87.01 ± 1.17	**87.13 ± 1.81**
4	76.25 ± 2.89	74.41 ± 1.29	84.59 ± 3.76	75.47 ± 2.47	89.26 ± 1.27	**89.33 ± 1.53**
5	77.57 ± 2.09	79.53 ± 1.79	87.87 ± 2.07	78.25 ± 1.75	90.10 ± 0.95	**90.34 ± 1.35**
6	79.22 ± 1.53	80.92 ± 1.56	88.08 ± 6.34	80.91 ± 2.02	91.42 ± 0.97	**91.80 ± 0.80**
7	82.45 ± 2.10	83.43 ± 1.63	92.00 ± 3.02	82.65 ± 1.92	91.72 ± 0.98	**92.09 ± 0.87**

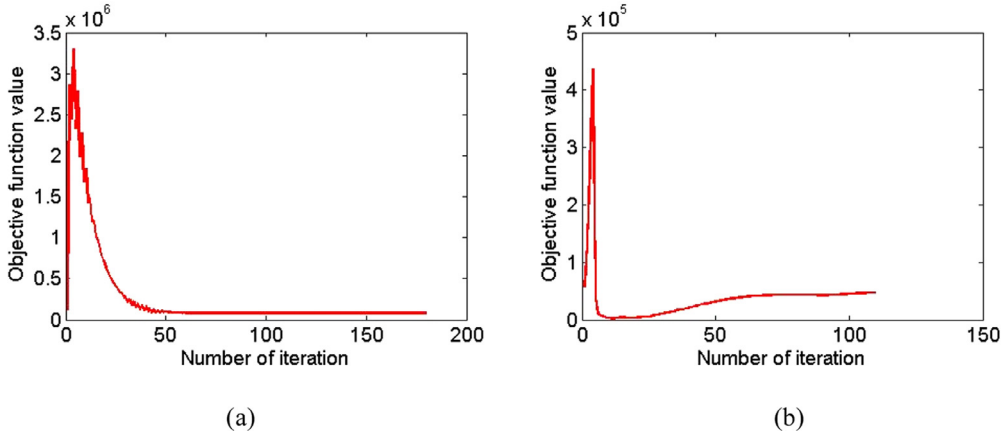

Fig. 2. Convergence curves of LRRADP. (a) and (b) are convergence curves of the LRRADP and LRRADP^2, respectively.

[17], Sparse Subspace Clustering (SSC) [35,36], and Robust LatLRR [18] methods. All algorithms first construct an affinity matrix by respective corresponding technique and then GFHF propagates the class labels from labeled samples to unlabeled samples based on the constructed affinity matrix/graph. In the classification procedure, we randomly select different numbers of samples per subject as labeled samples and use the remaining as unlabeled samples. All algorithms are run 10 times and then the mean classification results and standard deviation are reported and compared.

All algorithms are run on MATLAB 8.2.0 on a PC with double-core Intel(R) i5-3470 CPU at 3.2 GHz, RAM 8.00GB and Windows 7.0 operating system. It should be pointed out that the parameters of all these methods are carefully adjusted to obtain the best classification results. To improve the computation efficiency of the LRRADP, the feature dimensions of the data are reduced by using PCA to preserve 98% energy of the data.

5.1. Experiments on the COIL20 dataset

The COIL20 dataset (http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php) contains 1440 images of 20 objects and each object provides 72 images, which were captured from varying angles at pose intervals of five degree. The original images were normalized to 128×128 pixels and they are resized to a gray-scale image of 32×32 pixels for computational efficiency in our experiments, and then converted a 1024 dimensional scale-level feature for each image. Twelve examples selected from the COIL20 dataset are shown as in row 1 of Fig. 1. In this experiment, 1 to 7 images per subject are randomly selected as labeled samples and the rest are used as unlabeled samples, respectively. In each case, the mean classification accuracy and corresponding standard deviations are listed as in Table 1, where #Tr denotes the number of labeled samples of a subject. It can be seen that the proposed methods, including the LRRADP and LRRADP^2, outperforms other methods. Fig. 2 shows the convergence curves of the LRRADP and LRRADP^2 algorithms, respectively. We can see that both of LRRADP and LRRADP^2 have fast convergence speed.

5.2. Experiments on the Extended Yale B dataset

The Extended Yale B dataset (http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html) consists of 2432 human frontal face images of 38 subjects. Each subject contains about 64 images taken under different illuminations. A number of images in Extended Yale B dataset are seriously affected by shadows or reflection. The row 2 of Fig. 1 show some images of the same person from the Extended Yale B dataset. In the experiments, images selected from randomly 30 subject forms the test dataset, in which 1 to 7 images per subject are randomly selected as labeled samples and the remaining images are used as unlabeled samples. The classification results are elaborated as in Table 2, from which we can see that the LRRADP performs the best among all methods.

5.3. Experiments on the AR dataset

AR face dataset [37] is public available at http://www2.ece.ohio-state.edu/~aleix/ARdatabase.html. It contains over 4000 images corresponding to 126 persons (70 men and 56 women). These images were captured under different facial expressions, illuminations and occlusions, such as sunglasses and scarf. The images were taken under strictly controlled conditions. Each person participated in two sessions, separated by about two weeks. The same pictures were taken in both sessions. Some images of the same person from the AR face dataset are shown as in row 1 of Fig. 1. In our experiments, we generate a sub-dataset by using the images from first 30 subjects, each of which contains 26 images. Thus, there are 780 images in total are used in the sub-dataset. Among these images, 1 to 7 images per each subject are randomly labeled and the remaining images are used as testing samples. Table 3 summarizes the classification results obtained by using different methods. From the table we can see that, in most cases, the LR-

Table 2
The accuracy (%) of classification obtained using different methods on the Extended Yale B dataset.

#Tr	LRR	LatLRR	SSC	Robust LatLRR	LRRADP	LRRADP^2
1	38.28 ± 4.97	36.76 ± 2.93	47.48 ± 2.03	31.18 ± 8.03	**58.00 ± 1.93**	49.77 ± 3.28
2	44.61 ± 3.05	53.74 ± 1.55	64.18 ± 2.46	48.71 ± 6.89	**71.31 ± 2.77**	64.37 ± 1.29
3	57.62 ± 2.45	63.06 ± 2.81	72.16 ± 2.81	59.74 ± 5.35	**76.80 ± 2.87**	72.21 ± 0.97
4	65.22 ± 2.01	69.18 ± 2.46	76.30 ± 2.22	69.28 ± 2.59	**80.86 ± 1.62**	77.58 ± 1.73
5	70.22 ± 2.93	74.05 ± 2.13	79.88 ± 1.95	75.45 ± 2.97	**83.11 ± 1.80**	79.87 ± 1.39
6	72.61 ± 5.52	77.79 ± 2.12	82.27 ± 1.54	80.47 ± 1.65	**83.72 ± 1.34**	81.71 ± 1.24
7	76.13 ± 6.14	81.69 ± 1.82	84.14 ± 1.72	84.01 ± 1.28	**85.19 ± 1.60**	83.36 ± 0.74

Table 3
The accuracy (%) of classification obtained using different methods on the AR dataset.

#Tr	LRR	LatLRR	SSC	Robust LatLRR	LRRADP	LRRADP^2
1	40.53 ± 3.94	56.53 ± 3.65	50.29 ± 6.54	63.63 ± 2.15	**66.57 ± 3.95**	60.89 ± 3.25
2	68.78 ± 3.85	78.87 ± 2.62	71.22 ± 3.77	80.64 ± 1.24	**81.32 ± 3.43**	77.43 ± 3.16
3	82.57 ± 1.41	86.23 ± 1.88	80.12 ± 1.81	87.57 ± 0.98	**89.42 ± 1.83**	83.51 ± 1.55
4	87.27 ± 1.28	90.41 ± 1.34	86.29 ± 2.77	90.33 ± 1.26	**91.79 ± 1.60**	90.26 ± 0.94
5	91.11 ± 1.55	93.14 ± 1.51	88.70 ± 2.19	93.54 ± 0.80	**95.41 ± 0.87**	92.10 ± 1.71
6	92.13 ± 0.82	94.58 ± 1.08	91.45 ± 2.52	94.48 ± 0.98	**95.33 ± 1.13**	93.15 ± 1.15
7	94.11 ± 1.03	95.79 ± 0.65	94.05 ± 0.97	95.05 ± 1.48	**96.37 ± 1.24**	95.18 ± 1.36

Table 4
The accuracy (%) of classification obtained using different methods on the Isolet5 dataset.

#Tr	LRR	LatLRR	SSC	Robust LatLRR	LRRADP	LRRADP^2
1	11.60 ± 8.98	35.73 ± 2.72	19.98 ± 9.09	30.51 ± 4.01	**45.15 ± 2.71**	44.24 ± 1.91
2	30.08 ± 9.62	53.33 ± 2.10	45.02 ± 5.48	47.54 ± 2.80	**58.28 ± 1.57**	57.20 ± 1.75
3	45.44 ± 9.82	61.26 ± 3.40	56.06 ± 2.79	57.20 ± 1.51	64.64 ± 2.47	**65.01 ± 2.51**
4	65.24 ± 3.11	67.88 ± 1.25	63.40 ± 1.55	61.05 ± 1.65	68.29 ± 2.08	**69.22 ± 1.61**
5	70.35 ± 2.24	71.27 ± 1.22	66.17 ± 1.25	64.26 ± 1.38	72.31 ± 0.94	**72.46 ± 2.36**
6	74.21 ± 1.49	74.06 ± 1.65	67.78 ± 1.63	67.65 ± 1.13	74.16 ± 1.80	**74.77 ± 1.08**
7	76.60 ± 1.83	75.86 ± 2.18	70.60 ± 1.31	69.97 ± 1.32	75.77 ± 1.79	**77.25 ± 1.63**

RADP can achieve higher classification accurate rates than other graph construction methods.

5.4. Experiments on Isolet5 dataset

The Isolet5 dataset is available online at https://archive.ics.uci.edu/ml/datasets/ISOLET. In the generation of the Isolet5 dataset, 150 subjects spoke the name of each letter of the alphabet twice. Hence, 52 training examples from each speaker are obtained. The speakers are grouped into sets of 30 speakers each, and are referred to as Isolet1, Isolet2, Isolet3, Isolet4, and Isolet5. The data appears in Isolet $1+2+3+4$ data in sequential order and the Isolet5 is a separate file. In this experiments, the Isolet5 dataset is used, which consists of 26 alphabet voice data from 30 subjects, each of which provide twice voice. In other word, the Isolet5 contains 26 classes of voice data, each of which has about 60 samples. Specially, we note that the data of "m" is missing and it has 59 samples. Each data in Isolet5 dataset is normalized to a serial of 617 pixels. Table 4 lists the results of the classification by using different methods. From the table we can see that the proposed methods (LRRADP and LRRADP^2) can achieve higher classification accuracy than other benchmark methods. Moreover, the LRRADP^2 performs better than the LRRADP.

5.5. Experiments on MNIST dataset

The MNIST dataset (http://www.cad.zju.edu.cn/home/dengcai/Data/MLData.html) consists of more than seventy thousand handwritten images of 10 digits with sizes of 28×28 pixels. The forth row of Fig. 1 shows some typical images of the MNIST dataset. In the experiments, we use the first 4000 images of the MNIST dataset to form the test dataset, in which each digit has about 400 samples. We randomly select 30, 50, 80, 100, 150, 200 samples per digit as labeled samples and use the remaining images as unlabeled samples. The experimental results are detailed in Table 5. We can see from the table that the proposed LRRADP^2 performs the best.

5.6. Experiments on C-Cube dataset

The C-Cube dataset (http://ccc.idiap.ch) contains more than fifty thousand cursive characters extracted from cursive words, including both the upper and lower case of 26 letters [38,39]. The fifth row of Fig. 1 presents some typical examples of the C-Cube dataset. In the experiments, we randomly select 2000 characters of 'A-J' and 'a-j', each of which has 100 samples, to form the test dataset. All characters are in the center of the bitmaps with different sizes. To facilitate the experiments, we first reset the size of each bitmap to make it with the same length and width by filling "black" background, and keep the character in the center of the bitmap without changing the character's size. Then, we resize the bitmap into 30×30 pixels. Fig. 3 shows some normalized bitmaps. In classification experiments, 20, 30, 40, 50, 60 and 70 images per character are randomly selected as labeled samples and the remaining sam-

Fig. 3. Some normalized bitmaps of the C-Cube dataset.

Table 5
The accuracy (%) of classification obtained using different methods on the MNIST dataset.

#Tr	LRR	LatLRR	SSC	Robust LatLRR	LRRADP	LRRADP^2
20	31.30 ± 1.15	62.75 ± 0.82	72.05 ± 1.43	34.90 ± 3.73	70.33 ± 1.41	**74.71 ± 0.72**
50	47.12 ± 7.01	71.41 ± 1.01	81.18 ± 0.87	40.05 ± 7.40	77.52 ± 0.84	**83.09 ± 0.57**
80	66.83 ± 3.98	74.56 ± 0.54	84.19 ± 0.60	55.08 ± 12.36	81.87 ± 0.63	**86.12 ± 0.59**
100	74.56 ± 0.52	75.41 ± 0.75	85.18 ± 0.45	66.27 ± 12.94	82.92 ± 0.66	**87.45 ± 0.49**
150	77.97 ± 0.82	77.96 ± 0.61	87.01 ± 0.43	78.64 ± 0.89	85.12 ± 0.46	**89.46 ± 0.44**
200	79.60 ± 0.81	79.51 ± 0.75	87.76 ± 0.52	81.03 ± 0.67	86.36 ± 0.42	**90.32 ± 0.62**

Table 6
The accuracy (%) of classification obtained using different methods on the C-Cube dataset.

#Tr	LRR	LatLRR	SSC	Robust LatLRR	LRRADP	LRRADP^2
20	34.26 ± 9.26	42.01 ± 1.37	28.29 ± 2.47	27.21 ± 1.85	35.02 ± 2.14	**42.71 ± 1.23**
30	43.10 ± 0.85	45.86 ± 1.92	33.06 ± 2.81	30.89 ± 3.29	44.33 ± 1.53	**49.23 ± 1.46**
40	45.89 ± 0.84	48.57 ± 1.03	40.63 ± 5.09	33.72 ± 3.92	46.74 ± 1.36	**54.11 ± 1.78**
50	47.17 ± 1.70	49.93 ± 1.64	46.88 ± 6.49	35.65 ± 6.80	49.11 ± 1.23	**56.79 ± 1.37**
60	49.26 ± 1.35	50.65 ± 1.14	48.66 ± 5.40	42.13 ± 13.43	51.20 ± 1.38	**59.49 ± 1.52**
70	49.25 ± 1.62	51.37 ± 1.01	55.57 ± 7.79	45.63 ± 8.49	51.64 ± 1.42	**60.80 ± 1.65**

ples are used as unlabeled samples, respectively. The classification results are summarized in Table 6, from which we see that the LRRADP^2 achieve higher accuracy than the other methods.

5.7. Experiments on corrupted and noisy datasets

In order to evaluate the robustness of the proposed methods, we simulate contiguous occlusions and random pixel corruptions on various levels respectively. We select the first 15 persons from the Extended Yale B face dataset and generate four synthetic datasets by adding different levels of occlusions or noises, which are formed as follows. For the contiguous occlusions, the block occlusions are randomly added to different locations in original images of Extended Yale B dataset and the block sizes are 10×10 and 20×20, respectively. For the noisy synthetic datasets, we randomly add 10% and 20% "salt & pepper" noises on the original samples of the Extended Yale B dataset, respectively. Fig. 4 shows some examples from these four synthetic datasets. We test and compare the performance of different algorithms on these corrupted/noisy synthetic datasets. In each data set, different number of images

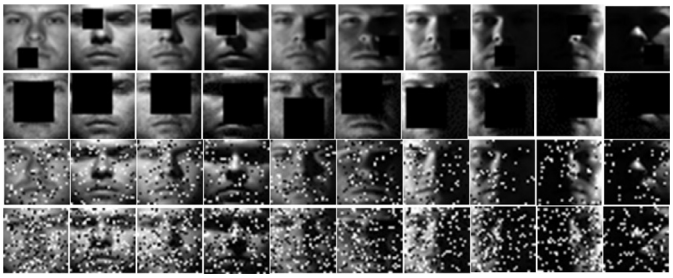

Fig. 4. Some examples of the corrupted/noisy images. The first and second rows are block corrupted images with size of 10×10 and 20×20 occlusions added on the original images of the Extended Yale B dataset, respectively; the third and fourth rows are noisy images with 10% and 20% "salt & pepper" noises added on the original images of the Extended Yale B dataset, respectively.

Table 7
The accuracy (%) of classification on the extended YaleB database with randomly block corruptions with size of 10×10 and 20×20, and with 10% and 20% "salt & pepper" noises, respectively.

	#Tr	LRR	LatLRR	SSC	Robust LatLRR	LRRADP	LRRADP^2
Corruptions (10×10)	5	50.68 ± 1.98	49.82 ± 1.16	48.00 ± 1.85	52.56 ± 1.49	59.45 ± 2.62	**60.56 ± 3.78**
	10	59.00 ± 2.17	59.42 ± 1.30	56.40 ± 1.23	61.97 ± 1.22	73.30 ± 2.60	**75.23 ± 1.78**
	15	65.77 ± 2.12	65.14 ± 2.02	60.50 ± 1.84	69.41 ± 2.22	79.67 ± 1.61	**82.52 ± 1.96**
	20	68.80 ± 2.37	68.98 ± 1.35	65.01 ± 1.31	73.52 ± 1.92	84.15 ± 1.49	**86.68 ± 0.83**
	25	70.22 ± 1.70	71.94 ± 1.45	68.07 ± 0.92	68.40 ± 1.90	85.18 ± 2.30	**88.37 ± 1.06**
	30	74.60 ± 1.53	73.07 ± 1.57	71.03 ± 2.30	79.13 ± 2.00	88.02 ± 1.21	**91.66 ± 1.68**
Corruptions (20×20)	5	22.91 ± 4.00	30.79 ± 1.84	32.83 ± 2.65	31.00 ± 2.46	35.30 ± 2.01	**44.38 ± 2.52**
	10	31.09 ± 7.69	38.75 ± 1.69	37.81 ± 2.82	38.85 ± 2.40	48.35 ± 2.54	**60.67 ± 1.92**
	15	37.96 ± 9.86	44.76 ± 1.35	42.35 ± 2.31	44.55 ± 1.88	58.66 ± 1.78	**72.10 ± 1.75**
	20	43.93 ± 7.11	58.05 ± 1.57	54.17 ± 1.54	58.77 ± 1.66	64.38 ± 1.64	**78.01 ± 1.38**
	25	51.93 ± 11.29	61.20 ± 1.41	61.36 ± 1.63	66.62 ± 2.65	68.80 ± 1.08	**83.59 ± 1.33**
	30	51.71 ± 11.56	62.32 ± 1.56	63.52 ± 2.98	70.32 ± 2.27	72.41 ± 2.18	**87.04 ± 2.20**
Noises (10%)	5	46.58 ± 2.03	47.19 ± 2.03	50.82 ± 1.73	50.63 ± 3.91	**57.25 ± 3.50**	56.63 ± 3.61
	10	48.50 ± 1.48	52.89 ± 1.85	55.33 ± 2.08	59.45 ± 2.10	70.65 ± 1.92	**72.39 ± 2.19**
	15	55.78 ± 1.57	59.03 ± 2.08	63.32 ± 1.26	69.08 ± 1.31	76.13 ± 1.00	**78.76 ± 2.05**
	20	56.44 ± 1.65	60.53 ± 1.98	65.70 ± 1.07	72.78 ± 2.32	79.94 ± 1.62	**80.81 ± 0.87**
	25	56.17 ± 1.53	61.78 ± 1.34	66.82 ± 1.05	74.48 ± 2.00	82.72 ± 1.40	**84.25 ± 1.39**
	30	56.83 ± 1.64	72.21 ± 2.15	67.72 ± 1.56	75.80 ± 2.01	84.53 ± 1.10	**85.63 ± 0.85**
Noises (20%)	5	23.16 ± 1.20	25.36 ± 1.87	24.45 ± 1.48	28.34 ± 5.55	30.87 ± 2.67	**38.30 ± 4.94**
	10	30.22 ± 1.55	34.10 ± 2.06	32.29 ± 0.80	41.74 ± 5.13	45.50 ± 2.10	**51.65 ± 2.47**
	15	36.03 ± 1.05	41.89 ± 1.33	40.05 ± 1.08	50.62 ± 1.31	54.44 ± 1.77	**59.40 ± 1.51**
	20	40.47 ± 0.80	46.42 ± 1.83	47.09 ± 1.00	51.89 ± 1.52	60.40 ± 1.27	**64.81 ± 1.65**
	25	40.76 ± 1.28	48.51 ± 1.41	46.19 ± 1.20	57.12 ± 1.38	62.36 ± 1.40	**67.70 ± 1.30**
	30	41.60 ± 1.78	48.31 ± 1.29	47.93 ± 1.46	58.80 ± 1.50	65.71 ± 2.40	**69.45 ± 1.37**

Table 8
The accuracy (%) of classification on the COIL20 database with randomly block corruptions with size of 10×10, and with 15% "salt & pepper" noises, respectively.

	#Tr	LRR	LatLRR	SSC	Robust LatLRR	LRRADP	LRRADP^2
Corruptions (15×15)	5	41.63 ± 2.10	55.82 ± 1.57	50.44 ± 3.20	43.05 ± 1.63	59.40 ± 1.30	**61.29 ± 1.74**
	10	58.88 ± 1.37	68.70 ± 1.87	61.14 ± 6.56	56.80 ± 1.38	71.95 ± 1.85	**74.79 ± 1.58**
	15	64.66 ± 1.59	73.25 ± 1.18	68.96 ± 8.07	65.11 ± 1.06	77.07 ± 0.96	**80.16 ± 1.01**
	20	67.06 ± 1.36	76.09 ± 1.45	77.24 ± 4.04	69.84 ± 0.96	81.81 ± 2.30	**83.72 ± 1.16**
	25	69.21 ± 2.17	78.24 ± 1.13	81.54 ± 2.14	74.05 ± 1.23	83.98 ± 0.96	**85.17 ± 1.31**
	30	70.26 ± 1.28	79.46 ± 1.49	84.20 ± 1.00	76.49 ± 1.89	85.93 ± 1.36	**87.05 ± 0.91**
Noises (15%)	5	69.10 ± 3.10	71.66 ± 1.87	68.35 ± 2.63	78.81 ± 1.74	77.96 ± 1.60	**80.20 ± 2.00**
	10	77.52 ± 1.92	80.78 ± 2.12	78.66 ± 1.85	83.52 ± 1.28	84.74 ± 1.42	**85.32 ± 1.13**
	15	80.69 ± 2.25	82.97 ± 1.36	81.43 ± 2.01	86.99 ± 1.14	87.34 ± 1.13	**87.44 ± 1.38**
	20	81.71 ± 1.54	86.19 ± 1.60	83.74 ± 0.97	87.97 ± 0.83	**89.88 ± 1.61**	88.38 ± 1.13
	25	84.43 ± 1.13	86.76 ± 1.14	85.01 ± 2.03	89.07 ± 0.89	**91.11 ± 0.81**	89.26 ± 1.41
	30	84.71 ± 0.83	87.45 ± 1.63	85.94 ± 1.12	89.96 ± 0.66	**92.33 ± 0.98**	89.56 ± 0.89

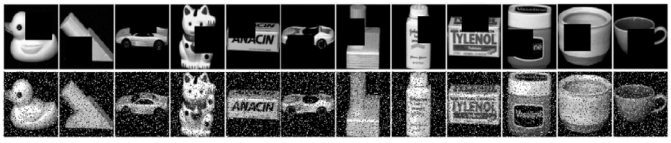

Fig. 5. Some examples of the corrupted/noisy images. The first row is block corrupted images with size of 15×15 occlusions added on the original images of the COIL20 dataset; the second row is noisy images with 15% "salt & pepper" noises added on the original images of the COIL20 dataset.

per person are randomly selected as labeled samples and the remaining images are used as unlabeled samples. The results on four types of synthetic datasets are shown as in Table 7. It is easy to see that the proposed methods consistently performs better than other methods on all four datasets.

Furthermore, we formed two datasets, including a corrupted dataset and a noisy dataset, by randomly adding the block occlusion with size of 15×15 and 15% "salt & pepper" noises on the COIL 20 dataset, respectively. Fig. 5 shows some examples from these two synthetic datasets. We evaluate the performances of the proposed methods as well as other algorithms on these corrupted and noisy datasets, in each of which different number of images per class are randomly selected as labeled samples and the rest images form the test set. The classification results on two synthetic datasets are summarized in Table 8, from which we can see that the LRRADP and LRRADP^2 methods consistently achieve higher classification accuracy than other methods on both COIL20 based synthetic datasets.

5.8. Experimental results analysis

Based on the evaluation results shown in the above tables, we have following findings. At first, experimental results on all datasets show that the LRRADP based methods, including the LRRADP and LRRADP^2 methods, can achieve higher classification accuracy than other methods in most conditions. It demonstrates that the LRRADP based methods can construct a discriminative affinity graph for the whole data. By using semi-supervised classification method, such as GFHF, the label information can be correctly propagated from the labeled to unlabeled samples over the graph.

Second, it is noticed that the LRRADP based methods can significantly increase the classification accuracy than other graph construction methods when there are only few labeled samples. For example, when only one sample per each subject is used as labeled sample, accuracies of the LRRADP based methods be 10% higher than that of the best among other methods for the COIL 20, Extended Yale B and Isolet5 datasets. The main reason is that the LRRADP effectively preserves the neighbor relationship among nearby samples, which is quite important for the semi-supervised label prediction. Therefore, the proposed methods are suitable to do the semi-supervised classification when only limited amount of samples are labeled.

Third, in general, as an improved LRRADP, the LRRADP^2 performs similarly with or better than the LRRADP method, which can be seen in the experimental results on the COIL20 and Isolet5 datasets. However, it is noticed that the LRRADP outperforms the LRRADP^2 for two face datasets. The possible reason is that most of images in these two face datasets have greater variations. For example, most of images in the Extended YaleB dataset are heavily shadowed and the images in the AR dataset were captured under different facial expressions and sunglasses or scarf occlusions, all of which lead to these images cannot be well projected into an appropriate subspaces. In other words, the LRRADP^2 possibly project these images into an unsuitable subspace resulting the drop of the accurate rate.

Forth, from the experimental results on the corrupted/noisy datasets, it is not hard to see that the proposed methods can significantly increase the classification accuracy on the block corrupted and random noisy datasets in most conditions. In other words, the adaptive distance penalty embedded in the LRR can effectively improve the robustness to the corruption and noise. Moreover, we can see that the LRRADP^2 performs much better than the LRRADP on both the real world handwritten digit/character and corrupted/noisy datasets. This is because that the Euclidean distance and linear combination in the LRRADP can be suitable adjusted to an appropriate representation. By appropriately assigning the projection matrix and then projecting the data set into the subspace, the LRRADP^2 is able to filter out some outlier and noise influence so as to construct a more discriminative affinity graph.

In summary, the LRRADP is suitable for the classification of those data with few noises and the LRRADP^2 is suitable for the real world handwritten and corrupted/noisy data.

6. Conclusions

In this paper, we considered the general problem of learning from labeled and unlabeled samples and classifying the unlabeled samples and proposed a novel low rank representation with adaptive distance penalty, named LRRADP, to learn the affinity graph of the data set. By embedding the adaptive distance penalty into the LRR, the obtained affinity graph can better not only capture the global clustering structure of the whole data but also preserve the local neighbor relationship of those data. Based on the affinity graph, the semi-supervised label propagation method, such as GFHF, can effectively propagate labels from the labeled samples to unlabeled samples. By projecting the data set into an appropri-

ate subspace, the LRRADP can be further improved to discover a more discriminative affinity graph. The improved LRRADP, named as LRRADP^2, shows competitive performance on the real world handwritten, block occlusions and random noises datasets. Experimental results on multiple datasets demonstrate the effectiveness of the proposed methods. Furthermore, the proposed methods are very effective and suitable for the semi-supervised classification when the labeled samples are relative limited.

Acknowledgment

This paper is partially supported by the National Natural Science Foundation of China (No. 61370163) and Shenzhen Municipal Science and Technology Innovation Council (No. JCYJ20130329154017293).

References

[1] J. Yan, M. Pollefeys, A general framework for motion segmentation: independent, articulated, rigid, non-rigid, degenerate and nondegenerate, in: ECCV, 2006, pp. 94–106.
[2] D. Cai, X. He, J. Han, Semi-supervised discriminant analysis, in: ECCV, 2007, pp. 1–7.
[3] X. Zhu, Semi-supervised learning literature survey, (2005), 1–59.
[4] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, S. Lin, Graph embedding and extensions: a general framework for dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell. 29 (1) (2007) 40–51.
[5] B. Ni, S. Yan, A. Kassim, Learning a propagable graph for semisupervised learning: classification and regression, IEEE Trans. Knowl. Data Eng. 24 (1) (2012) 114–126.
[6] F. Nie, S. Xiang, Y. Jia, C. Zhang, Semi-supervised orthogonal discriminant analysis via label propagation, Pattern Recogn. 42 (2009) 2615–2627.
[7] V. Sindhwani, P. Niyogi, M. Belkin, Linear manifold regularization for large scale semi-supervised learning, in: ICML, 2005, pp. 80–83.
[8] X. Zhu, Z. Ghahramani, J. Lafferty, Semi-supervised learning using Gaussian fields and harmonic functions, in: ICML, 2003, pp. 912–919.
[9] Y. Luo, D.C. Tao, B. Geng, C. Xu, S.J. Maybank, Manifold regularized multitask learning for semi-supervised multilabel image classification, IEEE Trans. Image Process. 22 (2) (2013) 523–536.
[10] F. Nie, D. Xu, I. Tsang, C. Zhang, Flexible manifold embedding: a framework for semi-supervised and unsupervised dimension reduction, IEEE Trans. Image Process. 19 (7) (2010) 1921–1932.
[11] R. He, W. Zheng, B. Hu, X. Kong, Nonnegative sparse coding for discriminative semi-supervised learning, in: CVPR, 2012, pp. 2849–2857.
[12] J. Wang, F. Wang, C. Zhang, H. Shen, L. Quan, Linear neighborhood propagation and its applications, IEEE Trans. Pattern Anal. Mach. Intell. 31 (9) (2009) 1600–1615.
[13] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, Y. Ma, Robust recovery of subspace structures by low-rank representation, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 171–184.
[14] G. Liu, Z. Lin, Y. Yu, Robust subspace segmentation by low-rank representation, in: ICML, 2010, pp. 663–670.
[15] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000) 888–905.
[16] K. Tang, R. Liu, Z. Su, J. Zhang, Structure-constrained low-rank representation, IEEE Trans. Neural Networks Learn. Syst. 25 (2) (2014) 2167–2179.
[17] J. Chen, J. Yang, Robust subspace segmentation via low-rank representation, IEEE Trans. Cybern. 44 (8) (2014) 1432–1445.
[18] G. Liu, S. Yan, Latent low-rank representation for subspace segmentation and feature extraction, in: ICCV, 2011, pp. 1615–1622.
[19] H. Zhang, Z. Lin, C. Zhang, J. Gao, Robust latent low rank representation for subspace clustering, Neurocomputing 145 (2014) 369–373.
[20] S. Wei, Z. Lin, Analysis and improvement of low rank representation for subspace segmentation, arXiv:1107.1561, 2011.
[21] E.J. Candes, X. Li, Y. Ma, J. Wright, Robust principal component analysis, J. ACM 58 (3) (2011) 11–47.
[22] L. Zhuang, S. Gao, J. Tang, J. Wang, Z. Lin, Constructing a non-negative low rank and sparse graph with data-adaptive feature, arXiv:1409.0964, 2014.
[23] L. Zhuang, H. Gao, Z. Lin, Y. Ma, X. Zhang, N. Yu, Non-negative low rank and sparse graph for semi-supervised learning, in: CVPR, 2012, pp. 2328–2335.
[24] X. Fang, Y. Xu, X. Li, Z. Lai, W.K. Wong, Learning a non-negative sparse graph for linear regression, IEEE Trans. Image Process. 24 (9) (2015) 2760–2771.
[25] X. Fang, Y. Xu, X. Li, Z. Lai, W.K. Wong, Robust semi-supervised subspace clustering via non-negative low-rank representation, IEEE Trans. Cybern. 46 (8) (2016) 1828–1838.
[26] F. Nie, X. Wang, H. Huang, Clustering and projected clustering with adaptive neighbors, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2014, pp. 977–986.
[27] J. Feng, Z. Lin, H. Xu, S. Yan, Robust subspace segmentation with block-diagonal prior, in: CVPR, 2014, pp. 1–8.
[28] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, B. Scholkopf, Learning with local and global consistency, Adv. Neural Inf. Process. Syst. 16 (16) (2014) 321–328.
[29] Z. Lin, R. Liu, Z. Su, Linearized alternating direction method with adaptive penalty for low-rank representation, in: NIPS, 2011, pp. 612–620.
[30] J. Yang, Y. Zhang, Alternating direction algorithms for l1-problems in compressive sensing, SIAM J. Scientific Comput. 33 (1) (2011) 250–278.
[31] Z. Lin, M. Chen, Y. Ma, The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices, in: NIPS, 2011, pp. 1–20.
[32] J. Yang, X. Yuan, Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization, Math. Comput. 82 (281) (2013) 301–329.
[33] J. Cai, E. Candes, Z. Shen, A singular value thresholding algorithm for matrix completion, SIAM J. Optim. 20 (4) (2008) 1956–1982.
[34] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM J. Imaging Sci. 2 (1) (2009) 183–202.
[35] E. Elhamifar, R. Vidal, Sparse subspace clustering, in: CVPR, 2009, pp. 2790–2797.
[36] E. Elhamifar, R. Vidal, Sparse subspace clustering: algorithm, theory, and applications, IEEE Trans. Pattern Anal. Machine Intell. 35 (11) (2013) 2865-2781.
[37] A.M. Martinez, R. Benavente, The AR Face Database, 1998 CVC Technical Report #24.
[38] L. He, H. Zhang, Iterative ensemble normalized cuts, Pattern Recogn. 52 (2016) 274–286.
[39] F. Camastra, M. Spinetti, A. Vinciarelli, Offline cursive character challenge: a new benchmark for machine learning and pattern recognition algorithms, in: ICPR, 2006, pp. 913–916.

Lunke Fei received the B.S. and M.S. degree in computer science and technology from East China Jiaotong University, China, in 2004 and 2007, He is currently pursuing the Ph.D. degree in computer science and technology at Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China. His current research interests include pattern recognition and biometrics.

Yong Xu received the B.S. and M.S. degrees in 1994 and 1997, respectively, and the Ph.D. degree in pattern recognition and intelligence system from the Nanjing University of Science and Technology, Nanjing, China, in 2005. Currently, he is with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China. His current research interests include pattern recognition, biometrics, bioinformatics, machine learning, image processing, and video analysis.

Xiaozhao Fang received the M.S. degree in computer science from Guangdong University of Technology in 2008, and the Ph.D. degree in computer science and technology at Shenzhen Graduate School, Harbin Institute of Technology in 2016. He has published more than 15 journal papers. His current research interests include pattern recognition and machine learning.

Jian Yang received the B.S. degree in mathematics from Xuzhou Normal University, in 1995, the M.S. degree in applied mathematics from Chang sha Railway University, in 1998, and the Ph.D. degree from the Nanjing University of Science and Technology (NUST), in 2002, with a focus on pattern recognition and intelligence systems. In 2003, he was a Post-Doctoral Researcher with the University of Zaragoza. He was a Post-Doctoral Fellow with the Biometrics Centre, the Hong Kong Poly technic University, from 2004 to 2006, and the Department of Computer Science, New Jersey Institute of Technology, from 2006 to 2007. He is currently a Professor with the School of Computer Scienceand Technology, NUST. His journal papers have been cited more than 1600 times in the ISI Web of Science, and 2800 times in Google Scholar. He has authored over 80 scientific papers in patternre cognition and computer vision. His research interests include pattern recognition, computer vision, and machine learning. He is currently an Associate Editor of Pattern Recognition Letters and the IEEE Transactions on Neural Networks and Learning Systems, respectively.