Knowledge-Based Systems 177 (2019) 82–93

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Data induced masking representation learning for face data analysis ☆

Tan Guo [a,*], Lei Zhang [b], Xiaoheng Tan [b], Liu Yang [a], Zhifang Liang [a]

[a] School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[b] School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China

ARTICLE INFO

Article history:
Received 3 August 2018
Received in revised form 10 April 2019
Accepted 11 April 2019
Available online 25 April 2019

Keywords:
Representation learning
Semi-supervised learning
Low-rank representation
Sparse representation
Subspace learning

ABSTRACT

Representation learning models, such as the sparse representation and low-rank representation, have shown pleasing efficacy in exploring the intrinsic data structures for pattern recognition tasks. However, conventional methods ignore the local geometric and similarity information among samples, and the performance is restricted. To address this issue, this paper proposes a novel **D**ata **I**nduced **M**asking **R**epresentation (DIMR) learning model by imposing explicit regularization and low-rank constraint. Specifically, DIMR is formulated for shrinking the representations of inter-class and non-neighbor samples. An extra representation regularization term is deployed with a data induced mask matrix, which can incorporate label and locality priors to guide the learning of affinity representation matrix. The affinity graph derived from DIMR is with low-rank, locality preservation (sparsity) and label guiding, such that it can better characterize the adjacent relationship between samples. Extensive experiments on benchmark face datasets demonstrate the superiority of DIMR for both semi-supervised classification and semi-supervised subspace learning tasks.

1. Introduction

With the rapid development of social network, such as Facebook, Instagram and Pinterest, large-scale image data is produced. How to organize and retrieve information from these image data attracts a great deal of attention from researchers in the field of computer vision, e.g., image classification and image understanding. Human face is probably the most popular image with rich pattern information, and therefore provides a good test bed for image modeling and analysis. In practice, one often needs to reveal the underlying data structures for many computer vision, pattern analysis and image processing applications [1]. Affinity graph is usually constructed based on the similarities between pairs of data points, and has been widely applied in data clustering, subspace learning and semi-supervised learning problems [2–6]. The basic idea of these methods is to learn an affinity graph to characterize the adjacent relationship between data points. Samples are the nodes of the graph, and the edges among these nodes that measure the similarity with respect to the sample pairs are generally defined as affinity. However, constructing a good affinity graph to accurately capture the underlying structure of the observed data is still a challenging problem [7].

Traditional methods, such as *k*-nearest neighbors and local linear reconstruction-based graph [8,9], mainly rely on pair-wise Euclidean distances in graphs construction. Therefore, they are sensitive to the choice of the neighborhood size that is used to compute the local information at each data point [3]. As suggested by Cheng et al. an informative graph should have three characteristics: discrimination, sparsity and adaptive neighbors [1]. Sparse graph derived from sparse representation (SR) has the advantages of robustness to noise, data-adaptive neighbors and sparsity [1,3,10,11], and therefore achieves a great success. However, there still some problems emerge. Firstly, traditional SR is unsupervised and has shown to have the problem of randomness [12–14]. That is, SR tends to randomly select a single representative sample from the high-correlation samples. Meanwhile, SR might select quite different samples to favor sparsity. Yu et al. [15] empirically observed that sparse results tend to be local, i.e., the non-zero coefficients are often assigned to samples nearby the represented data. Simultaneously, they further theoretically pointed out that under certain assumption, locality is more essential than sparsity, as locality must lead to sparsity but the reverse is not necessarily true. Therefore, a good graph should reveal the local neighborhood relationship of data points and preserve the intrinsic geometric structure of the manifolds [16]. Secondly, sparse graph can only reveal local geometry relationship and cannot characterize the global structure of data. Low rankness, as a global regularization, has been introduced

☆ No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.knosys.2019.04.006.

* Corresponding author.
E-mail address: guot@cqupt.edu.cn (T. Guo).

https://doi.org/10.1016/j.knosys.2019.04.006
0950-7051/

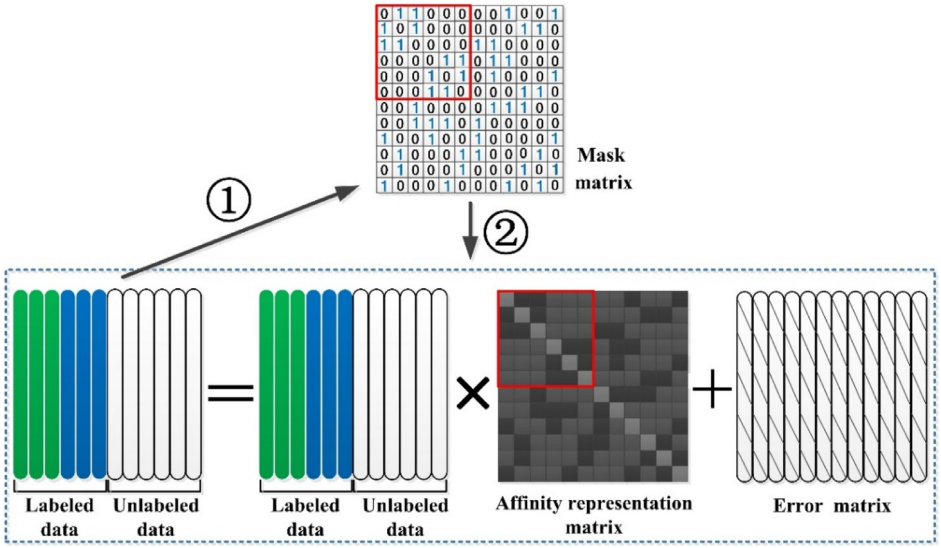


Fig. 1. An illustration for **D**ata **I**nduced **M**asking **R**epresentation (DIMR) learning. ① a mask matrix is firstly induced from the given dataset. The element w.r.t. the *i*th row and the *j*th column in the mask matrix codes the local geometric and similarity relationship of the *i*th and the *j*th sample in given dataset. ② with the guidance of mask matrix, DIMR model is then optimized to learn data-adaptive representation with local geometric structure and similarity information well preserved.

for visual analysis. Low-rank representation (LRR) is one of the representative methods for affinity graph construction [17–19]. To preserve the locality and similarity information in LRR learning, Yin et al. [20] developed a Laplacian regularized low-rank representation with a Laplacian regularization term. A locality-preserving low-rank representation (L^2R^2) is presented in [16], which constrains its linear representation to be nonzero only in a local neighborhood and simultaneously preserves the intrinsic geometric structure of the manifolds. Low-rank and sparse constraints have also been combined to reveal the local and global structures of data points [2].

In pattern analysis and signal processing community, semi-supervised learning has been attracting considerable attention over the past decades, because supervised learning needs to afford the expensive data labeling cost [21–23]. Recently, graph-based learning methods have been widely used to develop efficient algorithms for semi-supervised learning (SSL) tasks. Generally, most of these graph-based SSL algorithms adopt the so-called cluster assumption, i.e., neighbor points in the same low-dimensional smooth structure (e.g., cluster, subspace, or manifold) are likely to share the same label [2]. To facilitate the assumption, many methods approximate the underlying manifolds by constructing an undirected graph from the observed data points. In graph-based SSL methods, labeled and unlabeled samples are the nodes of a graph. The label information of the labeled samples can be propagated to the unlabeled ones over the graph through a regularized function on the graph [21,24]. Consequently, constructing a suitable graph that can well capture the intrinsic data structure is the key for graph-based SSL methods. However, the above-mentioned graph learning methods ignore precious label information in graph learning. Intuitively, it is beneficial to consider such information in the graph learning stage. Hence, Zhuang et al. proposed to construct a low-rank graph with label information considered [25], where the representation coefficients of samples from different classes are constrained to be zero. Although samples from different classes should have low similarity, the representation coefficients are not necessarily zeros.

Representation learning models have been proved to be efficient in revealing the complex intrinsic structures of a given dataset for pattern analysis. Parsimony of representation is an important criterion in representation learning. Sparse representation reveals the one-dimensional sparseness of data, while low-rank representation embodies the two-dimensional sparseness of data. Both of them have shown pleasing efficacy in exploring the low-dimensional structure of high-dimensional data. However, these methods cannot well enable the utilization of priors such as local and label information. Thus, the key problem for representation learning is to discover the low-dimensional structure of the observed data with prior information well preserved. We argue that for representation-based graph construction method, each sample should be represented mainly by its intra-class and neighbor samples. That is, the representation coefficients from its intra-class and neighbor samples should be larger than that of inter-class samples and non-neighbors. In view of this, we propose a novel Data Induced Masking Representation (DIMR) learning model by imposing an explicit regularization to address the above issue. In DIMR, a representation regularization term is deployed with a data induced mask matrix, which can incorporate label and locality information to guide the learning of affinity representation matrix. The implementation of DIMR is to enable the compactness of representation coefficients of the intra-class and neighbor samples, but shrink the representation coefficients of the inter-class and non-neighbor samples. DIMR inherits the merits of representation learning-based and traditional Euclidean distance-based graph construction models. Different from existing models, our model can learn informative and discriminant affinity representation matrix by simultaneously incorporating label and locality priors together with the low-rank constraint. As a by-product, the obtained affinity representation matrix tends to be sparse with the guidance of locality information. Fig. 1 shows an illustration for the proposed DIMR model. As described above, the graph derived from DIMR model have three main characteristics: low-rank, locality preservation (sparsity) and label guiding. In addition, the obtained representation matrix tends to be block-diagonal by taken into account the priors of both label and locality information. For clarity, we have highlighted the novelty and contribution of this paper as follows:

(1) We propose a novel **D**ata **I**nduced **M**asking **R**epresentation (DIMR) learning model by imposing explicit regularization with a data induced mask matrix for affinity representation

matrix learning. To the best of our knowledge, this is the first report on representational learning via data induced mask. DIMR can be easily extended for supervised or unsupervised learning tasks by incorporating different priori information induced mask matrix.

(2) An effective optimization strategy based on the alternating direction method of multipliers (ADMM) is devised to optimize the DIMR model, and the convergence property of the optimization algorithm is validated from both theoretical and experimental perspectives.

(3) The affinity graph derived from DIMR is imposed to be low-rank, locality preservation (sparsity) and label guiding, which can better characterize the adjacent relationship between samples. Experimental results show that the proposed DIMR model can achieve promising performance for both semi-supervised classification and semi-supervised subspace learning tasks on benchmark face datasets.

2. Related works

2.1. Affinity graph learning

Given observed dataset $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_l, \boldsymbol{x}_{l+1} \ldots \boldsymbol{x}_N] \in \Re^{d\times N}$, where $\boldsymbol{x}_i|_{i=1}^{l}$ and $\boldsymbol{x}_j|_{j=l+1}^{N}$ are labeled and unlabeled data points, respectively. Dataset $\mathbf{X}$ can be represented by a weighted undirected graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ associated with a weight matrix $\mathbf{W} = \{w_{ij}\}$. $\mathbf{V} = \{v_i\}_i^N$ is the vertex set. Each data point corresponds to a vertex of the graph, and $\mathbf{E} = \{e_{ij}\}$ is the edge set. Each edge e_{ij} connects nodes v_i and v_j with weight w_{ij}. The problem of graph construction is to determine the graph weight matrix $\mathbf{W}$. Classic graph learning methods, e.g. k-nearest-neighbor method or ϵ-ball based method, rely on pair-wise Euclidean distances, which needs manually set global parameter to determine the neighborhoods for each data point, and hence fail to generate datum-adaptive neighborhood [8,9]. Besides, Euclidean distances capture the local structures of data, and are sensitive to noise and errors [1]. As indicated by Cheng et al. sparse graph derived from sparse representation have the following three properties: robustness to data noise, sparsity and datum-adaptive neighborhood [1]. The neighborhood samples of a datum and the corresponding connecting edge weight are simultaneously obtained by solving an l_1-norm optimization problem, where each datum is represented as the linear combination of the remaining samples and noise term [10]. Sparse graph can only reveal the local structure of data and lacks a global structure constraint [3]. LRR adopts low rankness as the global constraint to discover the global structure of data [17–19]. Low-rank and sparse representation have been combined to simultaneously capture the global and local structures in data [2]. Locality information has also been considered for affinity graph learning [16].

2.2. Semi-supervised classification and subspace learning

For a dataset $\mathbf{X} \in \Re^{d\times N}$, its label matrix is denoted as $\mathbf{Y} = [\mathbf{Y}_l\ \mathbf{Y}_u] \in \Re^{N\times C}$. $\mathbf{Y}_l \in \Re^{l\times C}$ and $\mathbf{Y}_u \in \Re^{(N-l)\times C}$ are denoted as the label matrix for labeled and unlabeled data points respectively. Intuitively, similar data points in the same low-dimensional smooth structure (e.g. cluster assumption and manifold assumption) should have similar label. With this principle, GFHF (Gaussian field and harmonic function) [21] and LGC (local and global consistency) [24] have been proposed, which can learn a continuous classification function $\mathbf{F} = [\mathbf{F}_l\ \mathbf{F}_u] \in \Re^{N\times C}$ from label matrix $\mathbf{Y}_l \in \Re^{l\times C}$ and well-constructed graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ with both label fitness and manifold smoothness. GFHF and LGC are formulated as the following models (1) and (2), respectively.

$$\min_{\mathbf{F}\in\Re^{N\times C}} \frac{1}{2}\sum_{i,j=1}^{N} \left\|\mathbf{F}_{i:} - \mathbf{F}_{j:}\right\|_2^2 w_{ij} + \lambda_\infty \sum_{i=1}^{l} \|\mathbf{F}_{i:} - \mathbf{Y}_{i:}\|_2^2 \tag{1}$$

$$\min_{\mathbf{F}\in\Re^{N\times C}} \frac{1}{2}\sum_{i,j=1}^{N} \left\|\frac{\mathbf{F}_{i:}}{\sqrt{\mathbf{D}_{ii}}} - \frac{\mathbf{F}_{j:}}{\sqrt{\mathbf{D}_{jj}}}\right\|_2^2 w_{ij} + \lambda \sum_{i=1}^{l} \|\mathbf{F}_{i:} - \mathbf{Y}_{i:}\|_2^2 \tag{2}$$

where λ balances the label fitness and manifold smoothness, and λ_∞ means an extreme large penalty on $\sum_{i=1}^{l} \|\mathbf{F}_{i:} - \mathbf{Y}_{i:}\|_2^2$ for label fitness on labeled data. $\mathbf{F}_{i:}$ and $\mathbf{Y}_{i:}$ are the ith row of label matrix $\mathbf{F}$ and classification matrix $\mathbf{Y}$ for datum $\mathbf{x}_i$, respectively. The above models can be further written as

$$\min_{\mathbf{F}\in\Re^{N\times C}} tr\left(\mathbf{F}^{\mathrm{T}}\mathbf{L}\mathbf{F}\right) + tr\left(\mathbf{F}-\mathbf{Y}\right)^T \mathbf{U}_\infty \left(\mathbf{F}-\mathbf{Y}\right) \tag{3}$$

$$\min_{\mathbf{F}\in\Re^{N\times C}} tr\left(\mathbf{F}^{\mathrm{T}}\tilde{\mathbf{L}}\mathbf{F}\right) + tr\left(\mathbf{F}-\mathbf{Y}\right)^{\mathrm{T}} \mathbf{U}_\lambda \left(\mathbf{F}-\mathbf{Y}\right) \tag{4}$$

where $\tilde{\mathbf{L}}$ and $\mathbf{L}$ are normalized Laplacian matrix and Laplacian matrix of the weight matrix $\mathbf{W}$, respectively. $\mathbf{U}_\lambda$ is a diagonal matrix with the m elements as λ w.r.t. the labeled samples and with the rest $N-l$ diagonal elements as 0 w.r.t. unlabeled samples, respectively. $\mathbf{U}_\infty$ is also a diagonal matrix with the m elements as λ_∞ w.r.t. labeled samples and with the rest $N-l$ diagonal elements as 0 w.r.t. unlabeled samples, respectively.

Another extensively studied topic in SSL is semi-supervised subspace learning, e.g., Semi-supervised Discriminant Analysis (SDA) [26], in which the labeled data points were used to maximize the separability between different classes and the unlabeled data points were used to estimate the intrinsic geometric structure of the data. The performance of SDA heavily depends on the affinity graph [26].

3. Data induced masking representation learning model

3.1. Model formulation

Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2 \ldots \mathbf{x}_N] \in \Re^{d\times N}$ be a collection of N data points with dimension d. The representation coefficient matrix $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2 \ldots \mathbf{z}_N] \in \Re^{N\times N}$ of dataset $\mathbf{X}$ over a dictionary $\mathbf{A} \in \Re^{d\times N}$ is as follows:

$$\mathbf{X} = \mathbf{A}\mathbf{Z} \tag{5}$$

where each $\mathbf{z}_i$ is the representation coefficient of $\mathbf{x}_i$. Although the above problem can be solved as a least square optimization problem with accurate reconstruction, the approach ignores all structural information and trivial solutions are easily generated. Therefore, effective constraint on the coefficient matrix $\mathbf{Z}$ for revealing the intrinsic structure of data is expected. We pursue that when small neighborhoods of data are used in reconstructing each sample, only the most relevant data points are selected for each reconstruction. As a result, the data lying in a unique subspace tends to use the same group of data for reconstruction. Specifically, we address this problem by revealing the underlying structure of data and impose the *lowest rank* criteria. That is, we seek a low-rank representation matrix $\mathbf{Z}$ by solving the following problem.

$$\min_{\mathbf{Z}} \|\mathbf{Z}\|_* \quad \text{s.t.} \quad \mathbf{X} = \mathbf{A}\mathbf{Z} \tag{6}$$

As a popular heuristic to replace the rank function. $\|\cdot\|_*$ is the nuclear norm of a matrix, which can be computed as the sum of the singular values of the matrix. Given dataset matrix $\mathbf{X}$, we may use the dataset themselves as the dictionary, i.e., $\mathbf{A}$ can be simply chosen as $\mathbf{X}$ itself. In real applications, the data are often

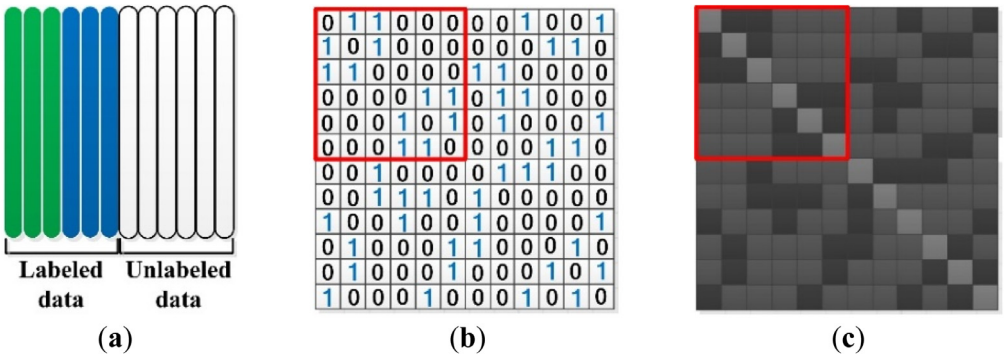


Fig. 2. An illustration for the inducing of mask matrix **M**. (a) is the given dataset composing of labeled and unlabeled data. (b) is the data induced mask matrix **M** according to the local and category relations between data points in (a). The ones in **M** mean that the associated samples are neighbors or have the same label. (c) is the desired affinity representation matrix. Note that the upper-left of (b) and (c) should have an approximate block-diagonal structure, which means larger intra-class similarity weights.

noisy and even grossly corrupted. Therefore, we use a sparse term to compensate the noise such that the negative effect can be reduced to some extent.

$$\min_{\mathbf{Z},\mathbf{E}} \|\mathbf{Z}\|_* + \lambda \|\mathbf{E}\|_1 \quad \text{s.t.} \quad \mathbf{X} = \mathbf{XZ} + \mathbf{E} \tag{7}$$

Since each data point is represented by other samples, a column $\mathbf{z}_i$ of $\mathbf{Z}$ can naturally characterize the contribution of other samples in reconstructing $\mathbf{x}_i$, and z_{ij} measures the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$. Among the solution of (7), some prior such as label or locality information is desired to be incorporated to guide the learning of $\mathbf{Z}$. As a result, we further impose regularization on $\mathbf{Z}$ to harness such prior information, which is formulated as follows:

$$\min_{\mathbf{Z},\mathbf{E}} \|\mathbf{Z}\|_* + \lambda \|\mathbf{E}\|_1 + \psi(\mathbf{Z}) \quad \text{s.t.} \quad \mathbf{X} = \mathbf{XZ} + \mathbf{E} \tag{8}$$

The regularization $\psi(\mathbf{Z})$ is expected to preserve the local geometric and similarity information of data points. Inspired by recent advances in representation learning models, in this paper, a regularization term is deployed through a data induced mask matrix, which can flexibly incorporate different kinds of prior information of data to guide the learning of affinity representation matrix. For semi-supervised learning, it is desired that if two samples are close in the intrinsic geometry of the data distribution or have the same label, they should have a large similarity coefficient. To this end, we propose to simultaneously introduce label and locality information together with low-rank constraint for affinity representation matrix learning. The representation coefficients on inter-class samples and non-neighbors are optimized to be small but are not necessarily be zeros. Formally, the proposed **D**ata **I**nduced **M**asking **R**epresentation (DIMR) learning model is formulated as

$$\min_{\mathbf{Z},\mathbf{E}} \|\mathbf{Z}\|_* + \lambda \|\mathbf{E}\|_1 + \frac{\alpha}{2} \|\mathbf{Z} - \mathbf{Z} \odot \mathbf{M}\|_{\mathrm{F}}^2 \quad \text{s.t.} \quad \mathbf{X} = \mathbf{XZ} + \mathbf{E} \tag{9}$$

where $\mathbf{M} \in \Re^{N \times N}$ is a data induced mask matrix with prior information of dataset $\mathbf{X} \in \Re^{d \times N}$ incorporated. $\odot$ is a Hadamard product operator of matrices. In DIMR, the mask matrix $\mathbf{M}$ should enable the utilization of label and locality prior efficiently. Mask matrix $\mathbf{M}$ is induced from dataset $\mathbf{X} \in \Re^{d \times N}$ as follows. For labeled data, if $l(\boldsymbol{x}_i) = l(\boldsymbol{x}_j)$, i.e., $\mathbf{x}_i$ and $\mathbf{x}_j$ have the same label, $\mathbf{M}_{ij} = 1$, $\mathbf{M}_{ji} = 1$, otherwise $\mathbf{M}_{ij} = 0$, $\mathbf{M}_{ij} = 0$. For unlabeled data, the value of $\mathbf{M}_{ij}$ is set as follows

$$\mathbf{M}_{ij}\left(\mathbf{M}_{ji}\right) = \begin{cases} 1, \textit{if } \mathbf{x}_i \in \mathrm{N}_k\left(\mathbf{x}_j\right) \text{ and } \mathbf{x}_j \in \mathrm{N}_k\left(\mathbf{x}_i\right) \\ 0, \qquad\qquad \text{otherwise} \end{cases} \tag{10}$$

where $\mathrm{N}_k\left(\boldsymbol{x}_j\right)$ denotes the set of k nearest neighbors of $\mathbf{x}_j$. With the above definition of $\mathbf{M}$, the term $\mathbf{Z} - \mathbf{Z} \odot \mathbf{M}$ can extract the representation coefficients on inter-class and non-neighbor samples. By optimizing $\|\mathbf{Z} - \mathbf{Z} \odot \mathbf{M}\|_F^2$ to be small, each labeled datum is promoted to be represented by its intra-class and nearby samples, and each unlabeled datum is encouraged to be represented by its neighbors. That is, the intra-class and neighbor data are encouraged to have large connecting weights. In this way, local geometric and similar information among data can be preserved in the obtained representation matrix $\mathbf{Z}$. An illustration for inducing the mask matrix $\mathbf{M}$ is shown in Fig. 2. Generally, the proposed model in Eq. (9) inherits the merits of Euclidean distance and representation-based graph leaning methods. The data neighbors and corresponding affinities are adaptively determined with pair-wise Euclidean distances and label prior considered.

3.2. Optimization for DIMR model

To optimize the proposed DIMR model (9), we first make an equivalent transformation by introducing one auxiliary variable to make the problem separable, and problem (9) can be rewritten as

$$\min_{\mathbf{Z},\mathbf{E},\mathbf{J}} \|\mathbf{J}\|_* + \lambda \|\mathbf{E}\|_1 + \frac{\alpha}{2} \|\mathbf{Z} - \mathbf{Z} \odot \mathbf{M}\|_{\mathrm{F}}^2 \tag{11}$$

$$\text{s.t. } \mathbf{X} = \mathbf{XZ} + \mathbf{E}, \quad \mathbf{Z} = \mathbf{J}$$

Then, we can get the following objective function of the problem using the inexact Augmented Lagrangian Multiplier (ALM) method [27] as follows.

$$\begin{aligned} \mathcal{L}(\mathbf{Z}, \mathbf{E}, \mathbf{J}, \mathbf{Y}_1, \mathbf{Y}_2) = {} & \|\mathbf{J}\|_* + \lambda \|\mathbf{E}\|_1 + \frac{\alpha}{2} \|\mathbf{Z} - \mathbf{Z} \odot \mathbf{M}\|_{\mathrm{F}}^2 \\ & + \langle \mathbf{Y}_1, \mathbf{X} - \mathbf{XZ} - \mathbf{E} \rangle \\ & + \langle \mathbf{Y}_2, \mathbf{Z} - \mathbf{J} \rangle + \frac{\mu}{2} \left(\|\mathbf{X} - \mathbf{XZ} - \mathbf{E}\|_{\mathrm{F}}^2 + \|\mathbf{Z} - \mathbf{J}\|_{\mathrm{F}}^2 \right) \end{aligned} \tag{12}$$

where $\mathbf{Y}_1$, $\mathbf{Y}_2$ are the Lagrangian multipliers, and $\mu > 0$ is a penalty parameter. The augmented Lagrangian is minimized along one coordinate direction at each iteration, i.e. minimizing the loss with respect to one variable with the other variables fixed. Specifically, we introduce the detailed optimization procedures in the $(k+1)$th iteration as follows.

Updating Z: Fix other variables and update $\mathbf{Z}$ by solving the following problem

$$\begin{aligned} \mathcal{L}(\mathbf{Z}) = {} & \frac{\alpha}{2} \|\mathbf{Z} - \mathbf{Z} \odot \mathbf{M}\|_{\mathrm{F}}^2 + \langle \mathbf{Y}_1, \mathbf{X} - \mathbf{XZ} - \mathbf{E} \rangle + \langle \mathbf{Y}_2, \mathbf{Z} - \mathbf{J} \rangle \\ & + \frac{\mu^k}{2} \left(\|\mathbf{X} - \mathbf{XZ} - \mathbf{E}\|_{\mathrm{F}}^2 + \|\mathbf{Z} - \mathbf{J}\|_{\mathrm{F}}^2 \right) \end{aligned} \tag{13}$$

$$\frac{\partial \mathcal{L}(\mathbf{Z})}{\partial \mathbf{Z}} = \alpha(\mathbf{Z} - \mathbf{Z} \odot \mathbf{M}) - \mathbf{X}^{\mathrm{T}}\mathbf{Y}_1 + \mathbf{Y}_2 + \mu^k\left(\mathbf{X}^{\mathrm{T}}(\mathbf{XZ} + \mathbf{E} - \mathbf{X}) + \mathbf{Z} - \mathbf{J}\right) \tag{14}$$

The closed-form solution can be obtained as follows by setting $\frac{\partial \mathcal{L}(\mathbf{Z})}{\partial \mathbf{Z}} = \mathbf{0}$.

$$\mathbf{Z}^{k+1} = \left((\alpha/\mu^k + 1)\mathbf{I} + \mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T(\mathbf{X} - \boldsymbol{E}) + \boldsymbol{J} + \left(\alpha\left(\mathbf{Z}^k \odot \mathbf{M}\right) + \mathbf{X}^T\boldsymbol{Y}_1 - \boldsymbol{Y}_2\right)/\mu^k\right) \tag{15}$$

Updating E: Fix other variables and update **E** by solving the following problem

$$\begin{aligned}\mathbf{E}^{k+1} &= \lambda\|\mathbf{E}\|_1 + \left\langle \mathbf{Y}_1^{\mathrm{T}}, \mathbf{X} - \mathbf{XZ}^{k+1} - \mathbf{E}\right\rangle + \frac{\mu^k}{2}\left\|\mathbf{X} - \mathbf{XZ}^{k+1} - \mathbf{E}\right\|_{\boldsymbol{F}}^2 \\ &= \frac{\lambda}{\mu^k}\|\mathbf{E}\|_1 + \frac{1}{2}\left\|\mathbf{E} - \left(\mathbf{X} - \mathbf{XZ}^{k+1} + \mathbf{Y}_1^{\mathrm{T}}/\mu^k\right)\right\|_F^2 \\ &= S_{\frac{\lambda}{\mu^k}}\left[\mathbf{X} - \mathbf{XZ}^{k+1} + \mathbf{Y}_1^{\mathrm{T}}/\mu^k\right]\end{aligned} \tag{16}$$

Updating J: Fix other variables and update **E** by solving the following problem

$$\begin{aligned}\mathbf{J}^{k+1} &= \underset{\mathbf{J}}{\operatorname{argmin}}\ \|\mathbf{J}\|_* + \langle \mathbf{Y}_2, \mathbf{Z} - \mathbf{J}\rangle + \frac{\mu^{\mathrm{k}}}{2}\|\mathbf{Z} - \mathbf{J}\|_{\mathrm{F}}^2 \\ &= \underset{\mathbf{J}}{\operatorname{argmin}}\ \|\mathbf{J}\|_* + \frac{\mu^{\mathrm{k}}}{2}\left\|\mathbf{J} - \left(\mathbf{Z} + \mathbf{Y}_2^{\mathrm{T}}/\mu^{\mathrm{k}}\right)\right\|_{\mathrm{F}}^2 = \mathbf{U}S_{1/\mu^k}[\Sigma]\mathbf{V}^{\mathrm{T}}\end{aligned} \tag{17}$$

where $(\mathbf{U}, \Sigma, \mathbf{V}^{\mathrm{T}}) = \mathrm{SVD}\left(\mathbf{Z} + \mathbf{Y}_2^{\mathrm{T}}/\mu^k\right)$ and $S_\varepsilon[\cdot]$ is the soft-thresholding (shrinkage) operator [27] defined as follows:

$$S_\varepsilon[x] = \begin{cases} x - \varepsilon, \textit{if } x > \varepsilon \\ x + \varepsilon, \textit{if } x < -\varepsilon \\ 0, \textit{otherwise} \end{cases} \tag{18}$$

Then, the multipliers and penalty parameter can be adjusted as follows:

$$\begin{cases} \mathbf{Y}_1^{k+1} = \mathbf{Y}_1^k + \mu^k\left(\mathbf{X} - \mathbf{XZ}^{k+1} - \mathbf{E}^{k+1}\right) \\ \mathbf{Y}_2^{k+1} = \mathbf{Y}_2^k + \mu^k\left(\mathbf{Z}^{k+1} - \mathbf{J}^{k+1}\right) \\ \mu^{k+1} = \min\left(\rho\mu^k, \mu_{max}\right) \end{cases} \tag{19}$$

Generally, the optimization process is outlined in Algorithm 1.

Algorithm 1. Solve DIMR model (11) using ADMM

Input: training dataset $\mathbf{X}$, balance parameters α and λ。

Initialization: $\mathbf{Z}^0 = \mathbf{J}^0 = \mathbf{E}^0 = \mathbf{0}$, $\mathbf{Y}_1^0 = \mathbf{Y}_2^0 = \mathbf{0}$, $\mu^0 = 10^{-5}$, $\mu_{max} = 10^8$, $\varepsilon = 10^{-6}$, $\rho = 1.1$.

1: While not convergence, **do**

2: Fix other variables, solve problem (15) to update $\mathbf{Z}^{k+1}$.

3: Fix other variables, solve problem (16) to update $\mathbf{E}^{k+1}$.

4: Fix other variables, solve problem (17) to update $\mathbf{J}^{k+1}$.

6: Update the multipliers and penalty parameters by (19).

7: Check the convergence conditions:

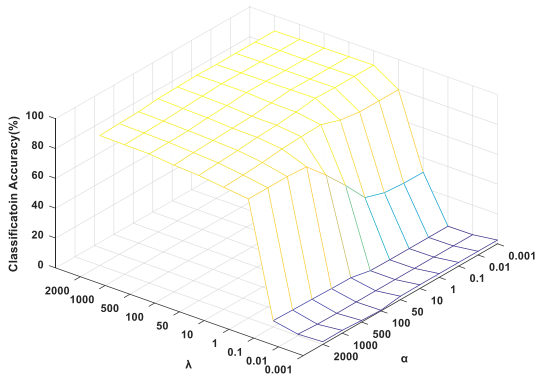
$$\left\|\mathbf{X} - \mathbf{XZ}^{k+1} - \mathbf{E}^{k+1}\right\|_\infty < \varepsilon, \left\|\mathbf{Z}^{k+1} - \mathbf{J}^{k+1}\right\|_\infty < \varepsilon$$

8: **End while**

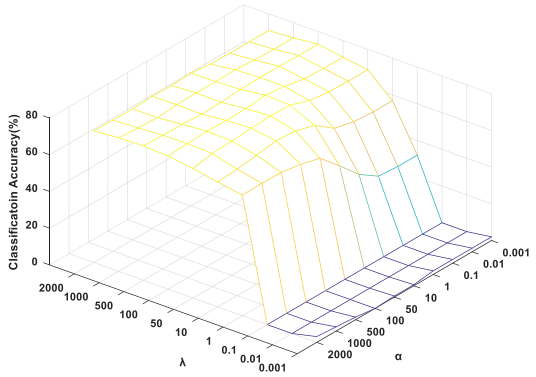
Output: **Z**

By solving the proposed DIMR model, the optimal representation matrix $\mathbf{Z}^*$ can be achieved. Then, we would construct an undirected graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$. The weight matrix and vertex set are denoted as $\mathbf{W} = \{\boldsymbol{w}_{ij}\}$ and $\mathbf{V} = \{v_i\}_{i=1}^N$, respectively. Each vertex v_i corresponds to one sample, and each edge in edge set $\mathbf{E} = \{\boldsymbol{e}_{ij}\}$ connects vertex v_i and v_j with a weight $\boldsymbol{w}_{ij}$. With the obtained representation matrix, z_{ij} (z_{ji}) can naturally measure the similarity between $\mathbf{x}_i$ $(\mathbf{x}_j)$ and $\mathbf{x}_j$ $(\mathbf{x}_i)$. The graph weight matrix $\mathbf{W}$ is defined as

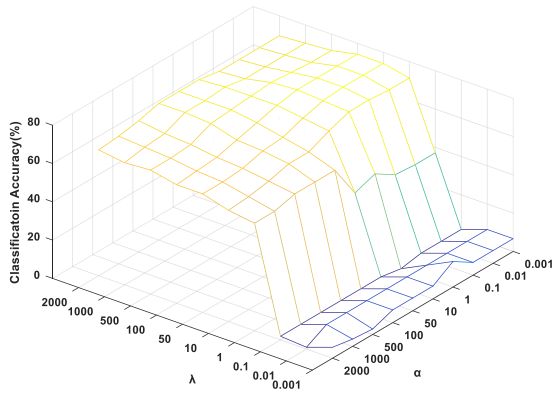
$$\mathbf{W} = \left(\mathbf{Z}^* + (\mathbf{Z}^*)^T\right)/2 \tag{20}$$



(a) ORL (#4)



(b) GT (#8)



(c) Yale (#4)

Fig. 3. Performance evaluation (%) of DIMR versus parameters α and λ on different datasets. (a) ORL database; (b) GT database; (c) Yale database.

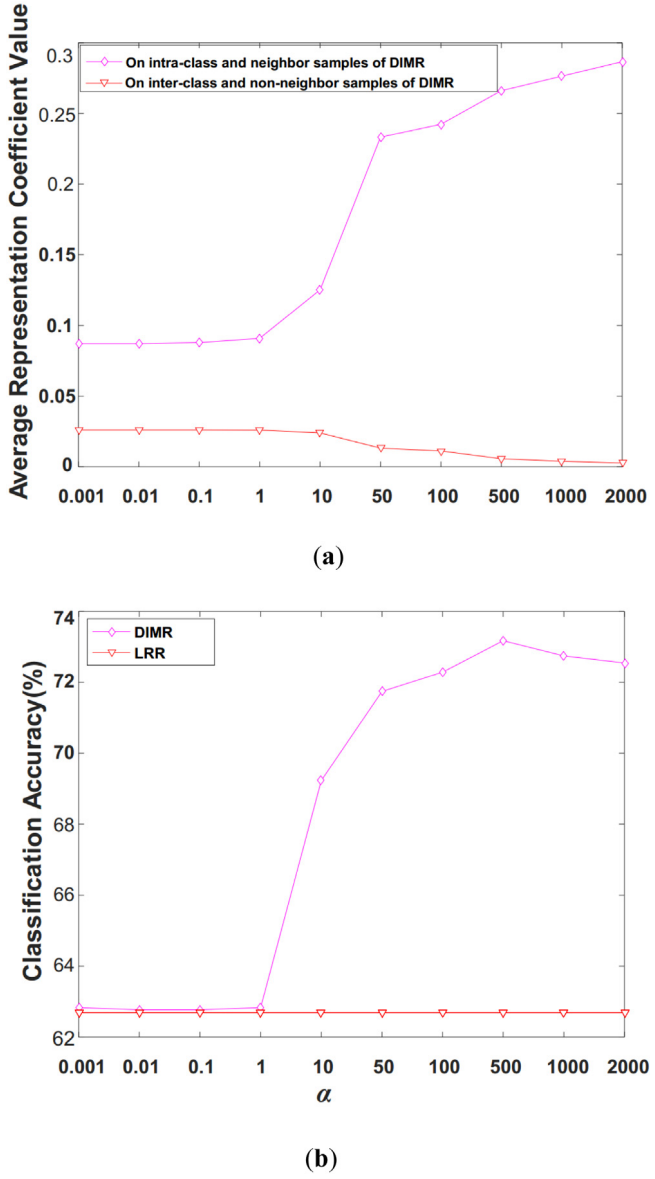


Fig. 4. (a) The average representation coefficient value of DIMR; (b) The semi-supervised classification accuracy of DIMR and LRR graphs corresponding to the different settings of parameter α in (a).

The method for constructing DIMR graph is summarized in Algorithm 2. Once the graph weight matrix is obtained, one can apply it for semi-supervised classification and semi-supervised discriminative subspace learning.

3.3. Complexity and convergence analysis

The major computational burden of Algorithm 1 lies in updating $\mathbf{Z}$ and $\mathbf{J}$ as they involve computation of matrix inverse and singular value. In Eq. (15), the matrix inversion operated for $\mathrm{N}\times\mathrm{N}$ matrix costs $\mathcal{O}\left(\mathrm{N}^3\right)$. In Eq. (17), SVD operated for matrix of $\mathrm{N}\times\mathrm{N}$ costs $\mathcal{O}\left(\mathrm{N}^3\right)$. The main computational complexity of Algorithm 1 is $\mathcal{O}\left(\tau\left(\mathrm{N}^3+\mathrm{N}^2 d\right)\right)$, where τ is the iteration number.

To solve the proposed DIMR model (9), an ADMM based iterative updating algorithm is developed. We will give the theoretical convergence of Algorithm 1. Classical ADMM aims to solve the following type of optimization problem,

$$\min_{\mathbf{z}\in\mathfrak{R}^n,\mathbf{g}\in\mathfrak{R}^m} f(\mathbf{z})+h(\mathbf{g}) \quad \text{s.t.} \quad \mathbf{Rz}+\mathbf{Tg}=\mathbf{u} \tag{21}$$

where $\mathbf{R}\in\mathfrak{R}^{p\times n}$, $\mathbf{T}\in\mathfrak{R}^{p\times m}$, $\mathbf{u}\in\mathfrak{R}^p$, f and h are convex functions. ADMM can be extended to solve matrix optimization problem as

$$\min_{\mathbf{Z}\in\mathfrak{R}^{n\times N},\mathbf{G}\in\mathfrak{R}^{m\times N}} f(\mathbf{Z})+h(\mathbf{G}) \quad \text{s.t.} \quad \mathbf{RZ}+\mathbf{TG}=\mathbf{U} \tag{22}$$

where $\mathbf{U}\in\mathfrak{R}^{p\times N}$, and the augmented Lagrange multiplier function is

$$\mathcal{L}_\mu(\mathbf{Z},\mathbf{G},\mathbf{C})=f(\mathbf{Z})+h(\mathbf{G})+\frac{\mu}{2}\|\mathbf{RZ}+\mathbf{TG}-\mathbf{U}\|_F^2 + \langle\mathbf{C},\mathbf{RZ}+\mathbf{TG}-\mathbf{U}\rangle \tag{23}$$

where $\mathbf{C}\in\mathfrak{R}^{p\times N}$ is Lagrange multiplier and $\mu>0$ is the penalty parameter. Optimization problem (11) is a special case of (22) by setting $\mathbf{R}=\begin{pmatrix}-\mathbf{I}_\mathrm{N}\\ \mathbf{X}\end{pmatrix}$, $\mathbf{T}=\begin{bmatrix}\mathbf{I}_\mathrm{N} & \\ & \mathbf{I}_\mathrm{D}\end{bmatrix}$, $\mathbf{G}=\begin{pmatrix}\mathbf{J}\\ \mathbf{E}\end{pmatrix}$, $\mathbf{U}=\begin{pmatrix}\mathbf{0}\\ \mathbf{X}\end{pmatrix}$. $\mathbf{I}_\mathrm{N}$ is an identity matrix with size of N. The constraint condition of DIMR can be formulated as the form of $\mathbf{RZ}+\mathbf{TG}=\mathbf{U}$. Then, the two primal variables in Eq. (23) can be solved alternatively and iteratively as follows:

$$\mathbf{Z}^{k+1}=\underset{\mathbf{Z}\in\mathfrak{R}^{n\times N}}{\operatorname{argmin}}\ \mathcal{L}_\mu\left(\mathbf{Z},\mathbf{G}^k,\mathbf{C}^k\right) \tag{24}$$

$$\mathbf{G}^{k+1}=\underset{\mathbf{G}\in\mathfrak{R}^{m\times N}}{\operatorname{argmin}}\ \mathcal{L}_\mu\left(\mathbf{Z}^{k+1},\mathbf{G},\mathbf{C}^k\right) \tag{25}$$

$$\mathbf{C}^{k+1}=\mathbf{C}^k+\mu\left(\mathbf{RZ}^{k+1}+\mathbf{TG}^{k+1}-\mathbf{U}\right) \tag{26}$$

Particularly, the optimization of $\mathbf{Z}$ in (24) is equivalent to optimize $\mathbf{Z}$ in Algorithm 1. Besides, the optimizations of $\mathbf{E}$ and $\mathbf{J}$ are independent with each other. Hence the optimizations of $\mathbf{E}$ and $\mathbf{J}$ can be accumulated in $\mathbf{G}$ using (25). In this way, solving DIMR model is an instance of classical ADMM problem. Algorithm 1 is equivalent to a two-block ADMM, and the global convergence is theoretically guaranteed [28–30].

4. Experimental evaluations

4.1. Parameters sensitivity and convergence analysis

There are two parameters α and λ in the objective function of the proposed DIMR model. This section will study the parameter sensitivity to the learnt graph for semi-supervised classification tasks. The semi-supervised classification accuracy under different parameter combinations with GFHF framework on three datasets is reported. α and λ are tuned in set {0.001,0.01,0.1,1,10,50,100, 500,1000,2000}. Three face datasets include Yale [31], ORL [32], and Georgia Tech (GT) [33] are employed in the experiments. From the experimental results in Fig. 3, where #Tr denotes the number of labeled samples per class selected for each dataset. We can see that the proposed model can achieve stable and promising performance under a large range of parameter settings. In detail, when α is very big ($>$500) and λ is small ($<$0.1), DIMR tends to achieve worse performance. Stable and encouraging performance can be achieved when α and λ are set with appropriate values, i.e., $500>\alpha>0.01$, $100>\lambda>50$. Thus, the parameter sets are suggested to get intra- and inter-class, neighbor and non-neighbor balanced low-rank representation.

Besides, the key idea of DIMR model is that each datum should be represented mainly by its intra-class and neighbor samples, so the representation coefficients among intra-class and neighbor samples should be large. On the contrary, the coefficients for inter-class and non-neighbors should be small. To verify this viewpoint, the mean values of representation coefficients for

Algorithm 2. DIMR graph learning

Input: training dataset $\mathbf{X}$, parameters α and λ. **Execute:**

1: Normalize all samples in the training dataset $\mathbf{X}$.

2: Run **Algorithm 1** to obtain the optimal affinity representation matrix $\mathbf{Z}$.

3: Normalize $\mathbf{Z}$ and construct the graph weight matrix $\mathbf{W}$ via

$$\mathbf{W} = \left(|\mathbf{Z}| + (|\mathbf{Z}|)^{\mathrm{T}}\right)/2$$

Output: graph weight matrix $\mathbf{W}$.

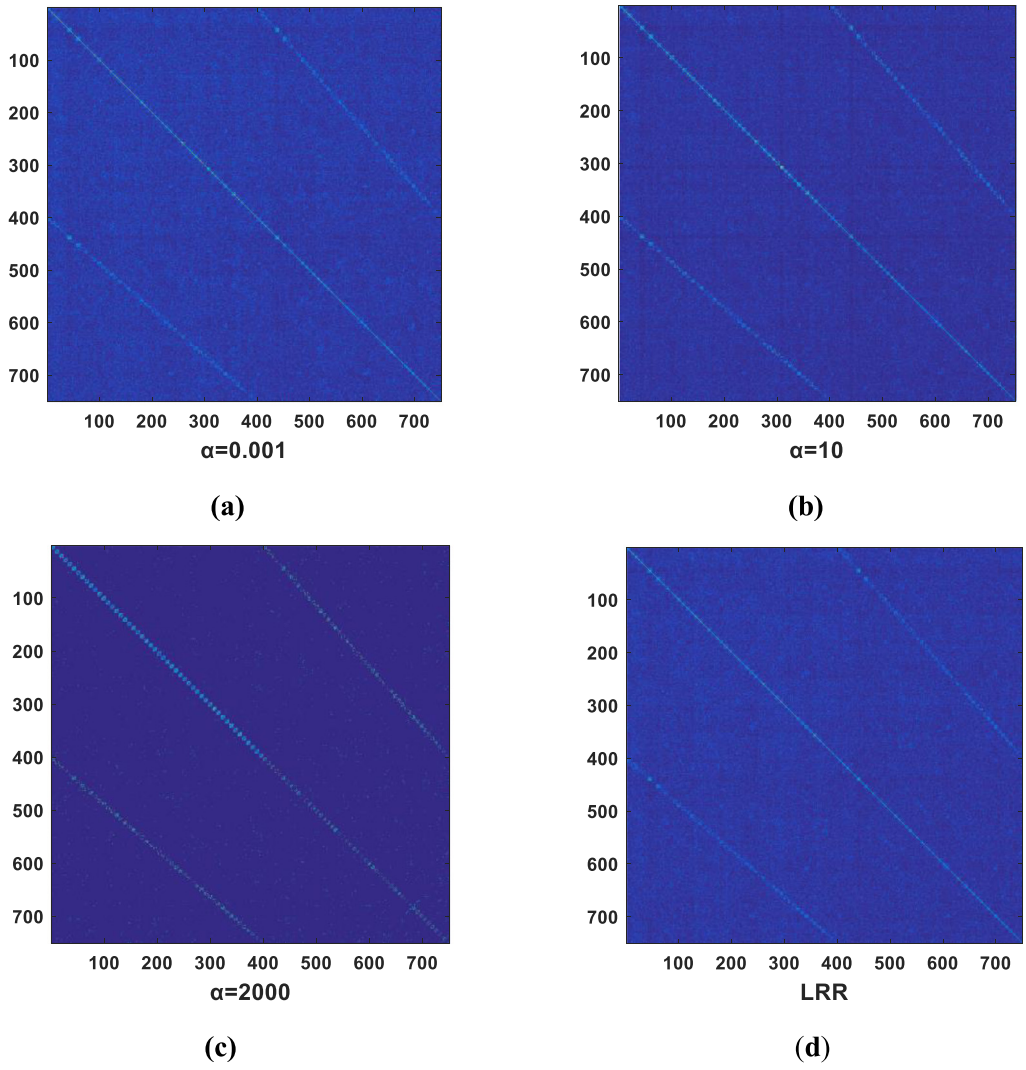
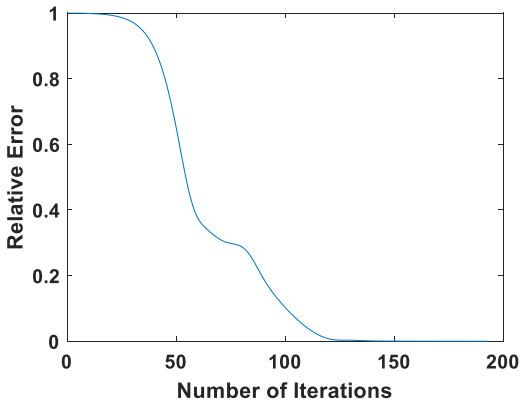


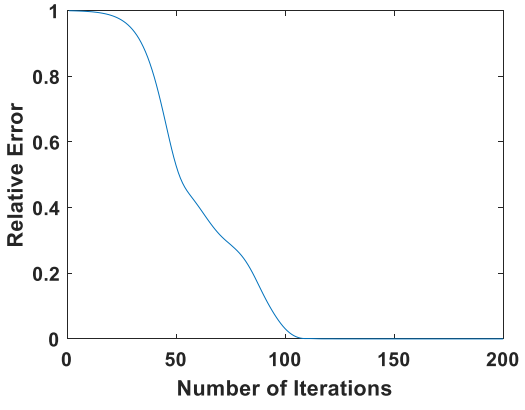
Fig. 5. The obtained DIMR graphs under different settings of α are visualized in (a), (b), and (c). LRR graph is shown in (d). In comparison with LRR graph, the developed DIMR graph becomes sparser with the enlargement of α, and can preserve the label and locality information.

intra-class and neighbor samples and the mean values of representation coefficients for inter-class and non-neighbor samples under different settings of parameter α are shown in Fig. 4(a). The classification accuracy w.r.t. the parameter is reported in Fig. 4(b). In addition, the obtained DIMR graphs are visualized in Fig. 5. From the results, one can observe that a larger regularization parameter α will result in the enhancement of the representation among intra-class and neighbor samples, and the representation among inter-class and non-neighbor samples shrink at the same time. Benefiting from the enhancement of representation among intra-class and neighbor samples, DIMR tends to achieve better semi-supervised classification performance than classic LRR. Figs. 4 and 5 show the reasonability and validity of the representation regularization introduced in DIMR.

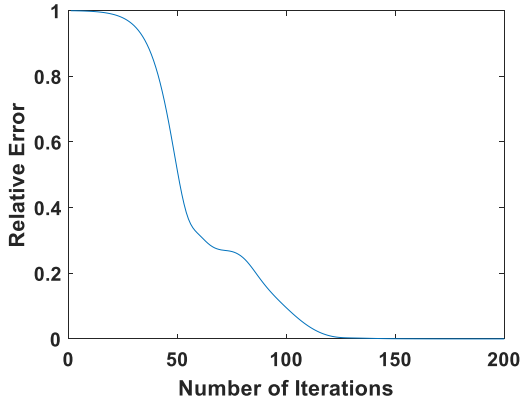
We have developed an ADMM based optimization Algorithm 1 to solve the DIMR model. The convergence property of Algorithm 1 on three datasets is presented in Fig. 6, where #Tr denotes the number of labeled training samples per class selected for each dataset. Similar to [34], the relative error (i.e., $\|\mathbf{X} - \mathbf{XZ} - \mathbf{E}\|_F / \|\mathbf{X}\|_F$) is employed to show the convergence. We can see that the relative error generally decreases with the increasing number of iterations, which shows the convergence property of Algorithm 1 from the experimental perspective.



(a) ORL (#4)



(b) GT (#8)



(c) Yale (#4)

Fig. 6. Convergence curves of the Algorithm 1 on different databases. (a) ORL database; (b) GT database; (c) Yale database.

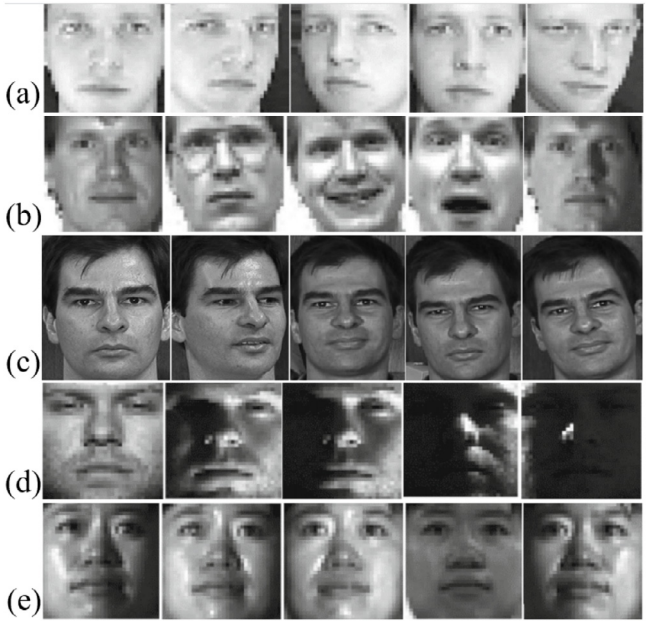


Fig. 7. Sample images used in our experiments. (a) ORL dataset; (b) Yale dataset; (c) GT dataset; (d) Extended Yale B dataset; (e) CMU PIE dataset.

Table 1
Description of the datasets used in the experiments.

Dataset	# Samples	# Dimension	# Classes
ORL [32]	400	1024	40
Yale [32]	165	1024	15
GT [33]	750	1024	50
Extended Yale B [31]	2414	1024	38
CMU PIE [35]	1680	1024	68

4.2. Experiment on semi-supervised classification scenario

This section will conduct more experiments to verify the performance of the proposed method together with GFHF framework for semi-supervised classification. Three datasets, i.e. ORL [32], Yale [32], and Georgia Tech (GT) [33] are employed. The description and some sample images of the datasets are shown in Table 1 and Fig. 7.

The ORL face database consists of 400 face images from 40 individuals with 10 images per person. The images were taken at different time, with lighting variation, facial expressions and facial details against a dark homogeneous background. In the experiments, each image in ORL database is manually cropped and resized to 32 × 32. A random subset with n (= 3, 4, 5, 6) images of each individual is selected for training and the rest for testing. For each n, we run the algorithm 10 times and the recognition rates as well as the standard deviations are reported in Table 2.

The Yale face database contains 165 gray scale images of 15 individuals, and each individual has 11 images. The images demonstrate variations in lighting condition and facial expression. In our experiments, each image in Yale database was manually cropped and resized to 32 × 32. A random subset with n (= 3, 4, 5, 6) images per individual is selected as labeled training data and the rest for testing. For each given n, we perform 10 times to randomly choose the training set and report the average recognition rates as well as the standard deviation in Table 2.

The GT face database contains 750 gray scale images of 50 individuals, and each individual has 15 images. The images demonstrate variations in pose, expression, illumination, and scale. In our experiments, a random subset with n (= 5, 8, 11) images per individual is selected as labeled training data and the rest is used for testing. For each given n, we perform 10 times to randomly choose the training set and report the average recognition rates as well as the standard deviation in Table 2.

Several state-of-the-art graph construction methods including KNN, LLE [8], SPG [36], SSC [3], LRR [17], NNLRS [2], and NSLLRR [20] are exploited for comparison. The number of neighbors in KNN, LLE, and the proposed DIMR is set as 5 and the Euclidean distance is adopted as the similarity measure. All the experimental results are reported in Table 2. From the results,

Fig. 8. Visualization of the first five basis vectors calculated by DIMR graph and SDA [26] on Extended Yale B database (a) and CMU PIE database (b).

Table 2
Experimental result on different databases using different methods.

Dataset	n	Comparing methods							Ours
		KNN	LLE [8]	SPG [36]	SSC [3]	LRR [17]	NNLRS [2]	NSLLRR [20]	DIMR
ORL	3	78.54 ± 2.02	83.75 ± 2.39	77.93 ± 2.73	81.11 ± 2.39	83.68 ± 2.65	67.43 ± 2.70	76.93 ± 3.11	**87.64 ± 2.06**
	4	79.88 ± 2.14	85.54 ± 2.99	81.17 ± 3.23	87.17 ± 3.19	85.63 ± 2.92	69.83 ± 2.74	80.00 ± 3.12	**90.67 ± 3.32**
	5	83.60 ± 2.07	89.30 ± 1.60	84.95 ± 2.30	92.35 ± 1.75	89.45 ± 2.35	71.50 ± 2.48	83.25 ± 1.64	**94.35 ± 1.20**
	6	83.56 ± 3.51	89.75 ± 3.06	79.63 ± 2.92	94.19 ± 1.93	90.88 ± 2.74	69.81 ± 4.90	83.88 ± 2.65	**95.44 ± 1.64**
YALE	3	49.83 ± 2.77	50.67 ± 2.91	53.25 ± 3.25	52.83 ± 2.19	58.67 ± 4.38	58.17 ± 4.25	57.58 ± 3.00	**64.00 ± 3.21**
	4	56.29 ± 2.96	58.29 ± 2.94	62.86 ± 5.33	57.71 ± 4.90	63.81 ± 4.16	58.10 ± 3.51	61.71 ± 4.23	**70.19 ± 4.58**
	5	57.78 ± 3.19	62.22 ± 5.37	67.00 ± 3.48	63.33 ± 4.54	68.44 ± 3.15	59.11 ± 3.81	65.33 ± 2.61	**75.44 ± 2.74**
	6	56.40 ± 3.56	61.60 ± 5.96	68.13 ± 6.42	67.73 ± 5.96	69.60 ± 6.31	59.73 ± 4.78	67.60 ± 5.22	**78.40 ± 5.21**
GT	5	63.68 ± 1.83	**68.60 ± 1.73**	58.38 ± 1.97	57.88 ± 1.51	54.70 ± 2.75	55.20 ± 2.52	57.74 ± 1.75	67.88 ± 1.30
	8	67.23 ± 1.90	74.34 ± 1.77	65.91 ± 2.81	70.57 ± 1.92	65.40 ± 1.76	63.29 ± 1.85	64.86 ± 1.69	**75.91 ± 1.29**
	11	70.55 ± 2.42	78.30 ± 3.22	70.90 ± 3.89	76.70 ± 3.23	68.90 ± 2.40	64.45 ± 2.57	67.45 ± 3.24	**81.25 ± 3.08**

we can observe that the proposed DIMR can achieve better performance by comparing with several state-of-the-arts. Comparatively speaking, the neighbors of samples in KNN graph cannot be adaptively determined but need manual setting, and also the structure of obtained graph is sensitive to noise. LLE tries to minimize the linear reconstruction error for graph construction. However, the minimization of linear reconstruction error in LLE is only processed within the sample neighbors defined by the k nearest neighbors. Thus, the structure of graph adjacency has been determined and LLE only generates the corresponding graph weights. In this way, the graph deduced by LLE is not optimal. SPG is formulated as a sparse coding problem with the non-negative constraint to adaptively determine the connecting neighbors and corresponding weights. SSC utilizes sparse constraint to construct the sparse graph, which can reveal the local geometry structure of data. However, Both SPG and SSC find the sparsest representation of each sample individually, which lacks of global constraints on their solutions. Therefore, these methods may be ineffective in capturing the global structures of data, which may reduce the performance when the data are grossly corrupted. LRR graph is constructed based on sample reconstruction principle, and adaptively determine the connecting relationship and weights between samples. Compared to the sparse graph, LRR often results in a dense graph, which is undesirable for graph-based SSL. NNLRS and NSLLRR only use the label information of the observed samples in the label propagation stage, while ignoring such valuable information during the graph learning, such that the adjacent relationship between samples cannot be well revealed. As mentioned above, the proposed semi-supervised classification methods can not only utilize the label information in label propagation stage, but also incorporate locality and label information in graph learning stage. With the low-rank constraint and explicit regularization for the affinity representation matrix, the obtained graph has the traits of low-rank, locality preservation (sparsity), and label guiding, which make the graph more discriminative and informative for semi-supervised classification task. As a result, the proposed method can alleviate the problems of the compared methods, and achieve higher classification accuracy.

To show the effectiveness of the proposed method, we perform statistical significance test to verify whether the improvement of DIMR over other methods is significant. More specifically, the popular t-test [37,38] was performed in a pair-wise manner on the null hypothesis that the improvement of DIMR over some competing method is insignificant. Two variables H and p are computed using t-test on the results from each pair of methods, where p denotes the probability of observing the given results. $H = 1$ denotes that the null hypothesis is rejected, and $H = 0$ denotes that the null hypothesis cannot be rejected, under some significance level α ($\alpha = 5\%$ or 1%). The test results are shown in Table 3, from which we can observe that $H = 1$ is achieved for all cases. The statistical significance of the proposed DIMR over other methods is clearly validated.

4.3. Experiments on semi-supervised subspace learning scenario

To evaluate the effectiveness of the proposed method on semi-supervised subspace learning task, we combine semi-supervised discriminant analysis (SDA) [36] with different graphs to compute the embedding. The compared methods include KNN, LLE [8], SSC [3], LRR [17], and NSLLRR [20]. The Extended Yale B [31] and CMU PIE [35] databases are adopted in the experiments. To make fair comparison, for all the evaluated algorithms we first apply PCA for data preprocessing by retaining 99% energy. The

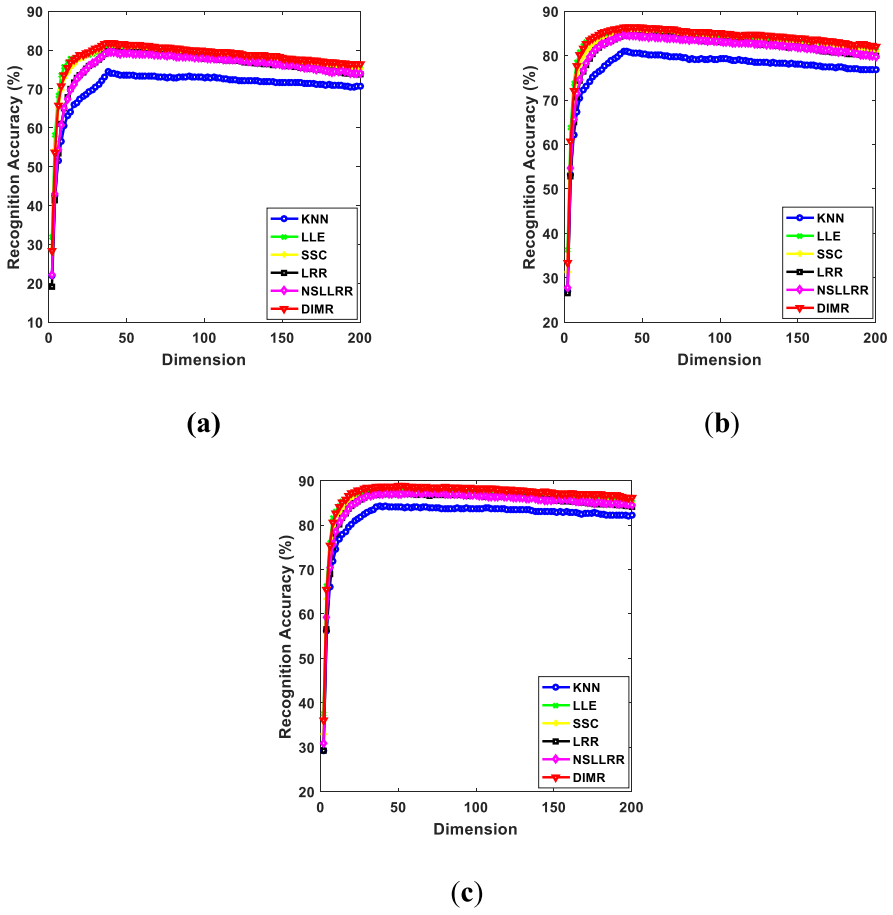


Fig. 9. Recognition rate versus dimension on the Extended Yale B database with 20 training samples of each individual. (a) 8 labeled training samples per person; (b) 10 labeled training samples per person; (c) 12 labeled training samples per person.

Table 3
Statistical hypothesis test by using t-test method of seven pairs of methods on semi-supervised classification task.

Pairs	⟨KNN, Ours⟩	⟨LLE, Ours⟩	⟨SPG, Ours⟩	⟨SSC, Ours⟩	⟨LRR, Ours⟩	⟨NNLRS, Ours⟩	⟨NSLLRR, Ours⟩
p	6.9×10^{-6}	1.9×10^{-3}	2.1×10^{-8}	1.9×10^{-4}	1.9×10^{-5}	1.9×10^{-6}	3.9×10^{-9}
$H(\alpha = 0.01)$	1	1	1	1	1	1	1
$H(\alpha = 0.05)$	1	1	1	1	1	1	1

Nearest Neighbor (NN) classifier is employed in the projected feature space for all the methods.

The Extended Yale B database consists of 2414 frontal face images of 38 individuals. Each individual contains about 64 images, taken under various laboratory-controlled lighting conditions. In the experiment, 20 samples per person are randomly selected as the training data, and the remaining are used for testing. For training samples, a random set of $n = 8, 10, 12$ of each individual are labeled and the rest are unlabeled. For each experimental setting, we run the algorithm code 10 times and report the recognition rates as well as the standard deviations with the increasing feature dimensionality. The CMU PIE dataset contains over 40,000 face images of 68 individuals. Images of each individual were acquired across 13 different poses under 43 different illumination conditions and 4 different expressions. Here we use a near frontal pose subset, namely C07, for experiments, which contains 1629 images of 68 individuals. Each individual has about 24 images. In the experiment, 10 samples of each individual are randomly selected as the training data, and the remaining are used for testing. For training samples, a random set of $n = 4, 5, 6$ samples per person are labeled and the rest are unlabeled. For each experimental setting, we also run the codes 10 times and report the average recognition rates as well as the standard deviations with the increasing feature dimensionality. The first five basis vectors calculated by DIMR graph and SDA on the two databases are visualized in Fig. 8. Besides, Figs. 9 and 10 plot the curves of average recognition accuracy versus the dimension on Extended Yale B and CMU PIE databases, respectively. Moreover, the details of experimental results, namely the maximum recognition rates and the standard deviations with different dimensionality based on different algorithms are summarized in Table 4. Benefiting from the discrimination of DIMR graph, SDA can achieve superior performance in comparison with related methods. The experiment shows the advantages of DIMR for semi-supervised subspace learning problem.

Similarly, we perform t-test [37,38] to show whether the improvement of DIMR over some competing method X is insignificant in a pairwise manner. The test results are shown in Table 5, from which one can see that the proposed DIMR method statistically outperforms other methods at the significance level $\alpha = 5\%$. The proposed DIMR also statistically outperforms others at the significance level $\alpha = 1\%$, except that the DIMR is not statistically significant over SSC graph.

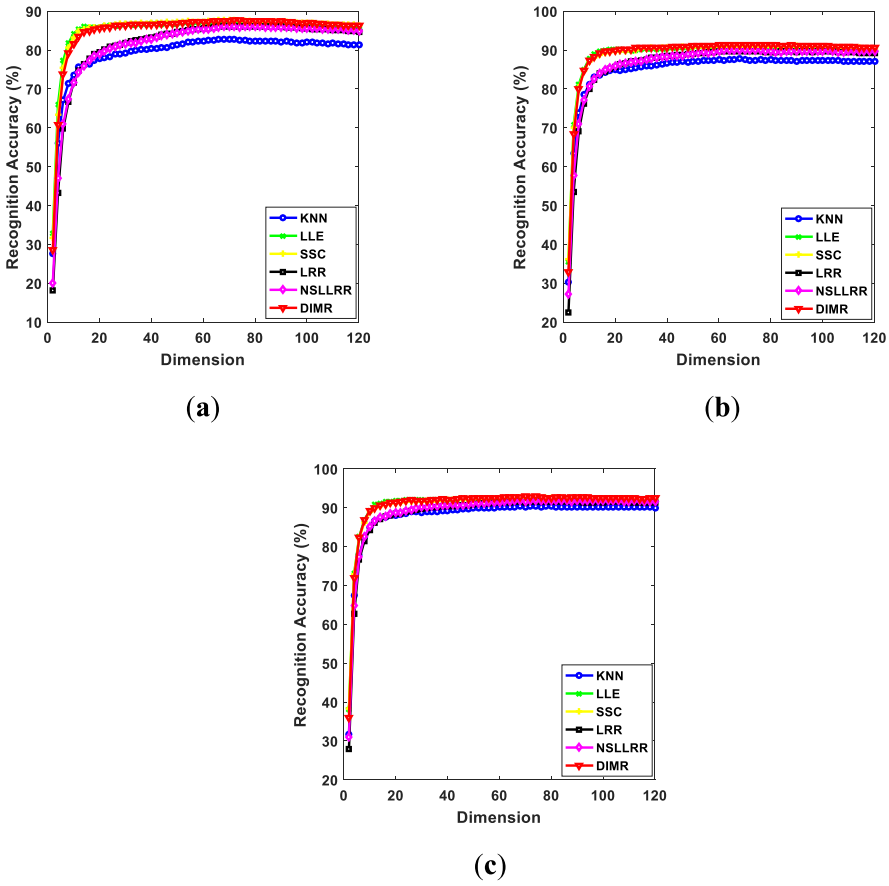


Fig. 10. Recognition rate versus dimension on the CMU PIE database with 10 training samples of each individual. (a) 4 labeled training samples per person; (b) 5 labeled training samples per person; (c) 6 labeled training samples per person.

Table 4
Recognition rates (%) and the corresponding standard deviations and dimensions (in parenthesis) on Extended Yale B and CMU PIE databases.

Dataset	n	Compared methods					Ours
		KNN graph	LLE graph	SSC graph	LRR graph	NSLLRR graph	DIMR graph
Extended Yale B	8	74.43 ± 1.44(38)	80.82 ± 1.86(38)	81.41 ± 1.45(38)	79.76 ± 2.04(44)	79.47 ± 1.55(38)	**81.91 ± 1.62(38)**
	10	81.04 ± 1.98(38)	85.42 ± 1.54(42)	85.75 ± 1.66(38)	84.66 ± 1.71(38)	84.67 ± 1.56(40)	**86.45 ± 1.71(42)**
	12	84.32 ± 1.28(38)	88.13 ± 1.11(38)	87.86 ± 0.95(38)	87.18 ± 1.24(56)	87.22 ± 1.12(66)	**88.80 ± 1.11(50)**
CMU PIE	4	82.74 ± 2.04(70)	87.27 ± 1.32(66)	**87.73 ± 1.13(66)**	86.17 ± 0.80(68)	86.15 ± 0.69(76)	87.69 ± 1.36(74)
	5	87.71 ± 1.61(68)	90.95 ± 0.94(78)	91.02 ± 0.81(90)	89.80 ± 1.04(102)	90.02 ± 1.11(68)	**91.41 ± 1.05**(76)
	6	90.34 ± 1.29(78)	92.41 ± 1.01(76)	92.52 ± 0.99(84)	91.68 ± 1.37(74)	91.88 ± 1.37(94)	**92.92** ± 0.93(70)

Table 5
Statistical hypothesis test by using t-test method of five pairs of methods on semi-supervised subspace learning task.

Pairs	⟨KNN graph, Ours⟩	⟨LLE graph, Ours⟩	⟨SSC graph, Ours⟩	⟨LRR graph, Ours⟩	⟨NSLLRR graph, Ours⟩
p	9.0×10^{-4}	2.2×10^{-3}	1.6×10^{-2}	4.1×10^{-5}	3.7×10^{-4}
$H(\alpha = 0.01)$	1	1	0	1	1
$H(\alpha = 0.05)$	1	1	1	1	1

5. Conclusions and further study

This paper presents a novel **D**ata **I**nduced **M**asking **R**epresentation (DIMR) learning model by explicitly regulating the affinity representation matrix with a data induced mask matrix. By coding both label and locality priors in the mask matrix, DIMR is formulated as an optimization problem of shrinking the representations between inter-class and non-neighbor samples with low-rank constraint. The obtained representation matrix is informative and discriminative, and shown to be low-rank and sparse with both label and local information preserved. The affinity graph derived from DIMR inherits the merits of Euclidean distance-based and representation-based graph leaning models. The proposed DIMR model is applied for semi-supervised classification and semi-supervised subspace learning tasks on benchmark face datasets, and the experimental results verify its effectiveness over many others. Future work will explore the application of the proposed model in a broader range of problems, such as supervised and unsupervised learning. Learning with hypergraph and deep representation will also be further investigated.

Acknowledgments

The authors would like to thank the Associate Editor and anonymous reviewers for their valuable comments, which greatly

improved the quality of this paper. This work was supported by National Natural Science Foundation of China (No. 61771079, 61571069, 61801072), the Science and Technology Research Program of Chongqing Municipal Education Commission (No. KJQN201800632, KJQN201800617), and Foundation and Frontier Research Project of Chongqing Municipal Science and Technology Commission (No. cstc2018jcyjAX0344, cstc2018jcyjAX0549, cstc2017zdcy-zdzxX0002).

Conflict of interest statement

We declare that we have no conflict of interest.

References

[1] B. Cheng, J. Yang, S. Yan, et al., Learning with l_1-graph for image analysis, IEEE Trans. Image Process. 19 (4) (2010) 858–866.
[2] L. Zhuang, S. Gao, J. Tang, et al., Constructing a nonnegative low-rank and sparse graph with data-adaptive features, IEEE Trans. Image Process. 24 (11) (2015) 3717–3728.
[3] E. Elhamifar, R. Vidal, Sparse subspace clustering: algorithm, theory, and applications, IEEE Trans. Pattern Anal. Mach. Intell. 35 (11) (2013) 2765–2781.
[4] W. He, H. Zhang, L. Zhang, W. Philips, W. Liao, Weighted sparse graph based dimensionality reduction for hyperspectral images, IEEE Geosci. Remote Sens. Lett. 13 (5) (2016) 686–690.
[5] Z. Kang, L. Wen, W. Chen, Z. Xu, Low-rank kernel learning for graph-based clustering, Knowl.-Based Syst. 163 (2019) 510–517.
[6] J. Chen, H. Mao, Y. Sang, Y. Zhang, Subspace clustering using a symmetric low-rank representation, Knowl.-Based Syst. 127 (2017) 46–57.
[7] X. Fang, Y. Xu, X. Li, Z. Lai, W.K. Wong, Learning a nonnegative sparse graph for linear regression, IEEE Trans. Image Process. 24 (9) (2015) 2760–2771.
[8] S. Yang, X. Wang, M. Wang, Y. Han, L. Jiao, Semi-supervised low-rank representation graph for pattern recognition, IET Image Process. 7 (2) (2013) 131–136.
[9] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.
[10] L. Qiao, S. Chen, X. Tan, Sparsity preserving projections with applications to face recognition, Pattern Recognit. 43 (1) (2010) 331–341.
[11] L. Zhang, W. Zuo, D. Zhang, LSDT: Latent sparse domain transfer learning for visual adaptation, IEEE Trans. Image Process. 25 (3) (2016) 1177–1191.
[12] A. Majumdar, R.K. Ward, Classification via group sparsity promoting regularization, in: ICASSP, 2009, pp. 861–864.
[13] J. Lai, X. Jiang, Class-wise sparse and collaborative patch representation for face recognition, IEEE Trans. Image Process. 25 (7) (2016) 3261–3272.
[14] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, J. Roy. Stat. Soc. B 68 (1) (2006) 49–67.
[15] K. Yu, T. Zhang, Y. Gong, Nonlinear learning using local coordinate coding, Adv. Neural Inf. Process. Syst. 22 (2009) 2223–2231.
[16] L. Zhuang, J. Wang, Z. Liu, A. Yang, Y. Ma, N. Yu, Locality preserving low-rank representation for graph construction from nonlinear manifolds, Neurocomputing 175 (2016) 715–722.
[17] G. Liu, Z. Lin, S. Yan, et al., Robust recovery of subspace structures by low-rank representation, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 171–184.
[18] G. Liu, Q. Liu, P. Li, Blessing of dimensionality: recovering mixture data via dictionary pursuit, IEEE Trans. Pattern Anal. Mach. Intell. 39 (1) (2016) 47–60.
[19] G. Liu, H. Xu, J. Tang, et al., A deterministic analysis for LRR, IEEE Trans. Pattern Anal. Mach. Intell. 38 (3) (2016) 417–430.
[20] M. Yin, J. Gao, Z. Lin, Laplacian Regularized low-rank representation and its applications, IEEE Trans. Pattern Anal. Mach. Intell. 38 (3) (2016) 504–517.
[21] X. Zhu, Z. Ghahramani, J. Lafferty, Semi-supervised learning using Gaussian fields and harmonic functions, in: Proc. 20th Int. Conf. Mach. Learn. (ICML), Vol. 20, 2003, pp. 912–919.
[22] T. Jebara, J. Wang, S.-F. Chang, Graph construction and b-matching for semi-supervised learning, in: Proc. 26th Annu. Int. Conf. Mach. Learn. (ICML), 2009, pp. 441–448.
[23] X. Zhu, Semi-Supervised Learning Literature Survey, Tech. Rep. 1530, Dept. Comput. Sci. Univ. Wisconsin-Madison, Madison, WI, USA, 2005.
[24] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, B. Schölkopf, Learning with local and global consistency, in: Advances in Neural Information Processing Systems, Vol. 16, MIT Press, Cambridge, MA, USA, 2003, pp. 595–602.
[25] L. Zhuang, Z. Zhou, S. Gao, J. Yin, Z. Lin, Y. Ma, Label information guided graph construction for semi-supervised learning, IEEE Trans. Image Process. 26 (9) (2017) 4182–4192.
[26] D. Cai, X. He, J. Han, Semi-supervised discriminant analysis, in: Proc. 11th ICCV, 2007, pp. 1–7.
[27] Z. Lin, R. Liu, Z. Su, Linearized alternating direction method with adaptive penalty for low rank representation, Adv. Neural Inf. Process. Syst. (NIPS) (2011) 612–620.
[28] R. Glowinski, P. Le Tallec, Augmented Lagrangian and Operator Splitting Methods in Nonlinear Mechanics, SIAM, Philadelphia, PA, USA, 1989.
[29] E. Esser, Applications of Lagrangian-based alternating direction methods and connections to split bregman, CAM Rep. 9 (2009) 31.
[30] J. Eckstein, D.P. Bertsekas, On the douglas-rachford splitting method and the proximal point algorithm for maximal monotone operators, Math. Program. 55 (1) (1992) 293–318.
[31] A.S. Georghiades, P.N. Belhumeur, D.J. Kriegman, From few to many: Illumination cone models for face recognition under variable lighting and pose, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 643–660.
[32] D. Cai, X. He, J. Han, H.J. Zhang, Orthogonal laplacianfaces for face recognition, IEEE Trans. Image Process. 15 (11) (2006) 3608–3614.
[33] Ara V. Nefian, Monson H. Hayes, Maximum likelihood training of the embedded HMM for face detection and recognition, in: Proc. Int'L Conf. Image Processing, Vol. 1, 2000, pp. 33–36.
[34] S. Li, Y. Fu, Learning robust and discriminative subspace with low-rank constraints, IEEE Trans. Neural Netw. Learn. Syst. 27 (11) (2016) 2160–2173.
[35] T. Sim, S. Baker, M. Bsat, The CMU pose, illumination, and expression database, IEEE Trans. Pattern Anal. Mach. Intell. 25 (12) (2003) 1615–1618.
[36] R. He, W.S. Zheng, B.G. Hu, X.W. Kong, Non-negative sparse coding for discriminative semi-supervised learning, in: Proc. CVPR, Providence, RI, USA, 2011, pp. 2849–2856.
[37] R.R. Wilcox, Introduction to Robust Estimation and Hypothesis Testing, second ed., Elsevier Academic Press, 2005.
[38] Demišar Janez, D. Schuurmans, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (1) (2006) 1–30.