Adaptive Local Embedding Learning for Semi-Supervised Dimensionality Reduction

Feiping Nie, Zheng Wang, Rong Wang, *Member, IEEE*, and Xuelong Li, *Fellow, IEEE*

Abstract—Semi-supervised learning as one of most attractive problems in machine learning research field has aroused broad attentions in recent years. In this paper, we propose a novel locality preserved dimensionality reduction framework, named Semi-supervised Adaptive Local Embedding learning (SALE), which learns a local discriminative embedding by constructing a k_1 Nearest Neighbors (k_1NN) graph on labeled data, so as to explore the intrinsic structure, i.e., sub-manifolds from non-Gaussian labeled data. Then, mapping all samples into learned embedding and constructing another k_2NN graph on all embedded data to explore the global structure of all samples. Therefore, the unlabeled data and their corresponding labeled neighbors can be clustered into same sub-manifold, so as to improve the discriminative power of embedded data. Furthermore, we propose two semi-supervised dimensionality reduction methods with orthogonal and whitening constraints based on proposed SALE framework. An efficient alternatively iterative optimization algorithm is developed to solve the NP-hard problem in our models. Extensive experiments conducted on several synthetic and real-world data sets demonstrate the superiorities of our methods on local structure exploration and classification task.

Index Terms—Semi-supervised dimensionality reduction, local embedding learning, adaptive neighbors, graph-based model

1 Introduction

In the epoch of Artificial Intelligence, machine learning achieves excellent performance in many applications such as multimedia understanding [1], crowd analysis [2], etc. Semi-supervised learning [3], [4] has been extensively studied among the machine learning researchers over the past several decades. One key reason is that it is easy to acquire a large collection of unlabeled data while data labeling is also time-consuming in many real world applications. Aa a consequence, the need for large scale labeled data sets is a major obstacle to the applicability of supervised learning. Therefore, it is inevitable to find an efficient way to learn more discriminative information from limited labeled data and large amount of unlabeled data. Semi-supervised learning is a successful technique that focuses on extracting discriminative information from both labeled and unlabeled data effectively. Co-training [5] and transductive SVM [6] are the two classical methods in early works [7].

Specifically, there is an assumption that data points with small distance should be shared with same label information. As a result, many graph-based methods have been proposed whose core idea is to learn the structure of data via graph in which each vertex is a sample and a small distance between two samples should be assigned a large weight. Among them, label propagation [8], Gaussian field harmonic function [9], local and global consistency [10] and random walk [11] are the most pioneered transductive methods. Although these methods acquire some certain results, transductive model always suffers from the issue called out-of-sample problem [12], namely, they are lack of the ability of dealing with unseen data sets. So this kind of methods have to put the new data into training set and retrain the model again and again, which is time-consuming.

To address above issues, Belkin *et al.*, first propose a semi-supervised framework that incorporates labeled and unlabeled data in a general-purpose learner [13], such that several existing supervised algorithms have been extended to semi-supervised learning, including Laplacian regularized least squares (LapRLS) and Laplacian support vector machines (LapSVM) by using the geometrical regularization term. Additionally, some specific paradigms can also be extended to semi-supervised learning such as Graph Regularized Non-negative Matrix Factorization (GRNMF) [14], Graph Regularized Sparse Coding (GRSC) [15], Laplacian Regularized Low-Rank Representation (LRLRR) [16], etc.

Recent successful studies in manifold learning [17] and Graph Embedding (GE) [18] have verified that the intrinsic structure of high-dimensional data essentially hidden at the low-dimensional subspace. GE-based algorithms aim to exploit the underlying sub-manifold structure of high-dimensional data by constructing a graph on all samples. According to these facts, several supervised dimensionality reduction approaches based on GE technique have been proposed, such as LLE [19], Laplacian Eigenmap [20], LFDA [21], LADA [22], NMMP [23] and LSDA [24].

Furthermore, GE technique has been widely applied in semi-supervised dimensionality reduction as well, and a large family of GE-based semi-supervised methods consist of two terms i.e., the label fitness term and manifold regularization. The former is supervised part that aims to integrate the label information into model and transforms original data into low-dimensional subspace. Meanwhile

- *The authors are with the School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, Shanxi 710072, P.R. China. E-mail: {feipingnie, zhengwangml}@gmail.com, wangrong07@tsinghua.org.cn, li@nwpu.edu.cn.*

Manuscript received 16 December 2018; revised 6 September 2020; accepted 16 December 2020. Date of publication 5 January 2021; date of current version 12 September 2022.
(Corresponding author: Feiping Nie.)
Recommended for acceptance by N. Chawla.
Digital Object Identifier no. 10.1109/TKDE.2021.3049371

the latter focuses on smoothing the manifold among all points, so as to explore the manifold structure of unlabeled data samples. Among these methods, the label fitness term usually adopts Least Squares Regression model (LSR). For example, originally, Zhang *et al.*, [25] exploit a regression residue term to instead label fitness term for binary classification. Moreover, Nie *et al.*, [26] propose an unified framework, named Flexible Manifold Embedding (FME) that relaxes the hard linear constraint in [25], which is helpful to handle the data sampled from a nonlinear manifold. Semi-supervised Elastic Embedding (SEE) [27] relaxes the rigid linear embedding constraint and uses an elastic embedding constraint on the predicted label matrix for exploring the manifold structure. It is evident that the noisy data points nearby the decision boundary have negative effects on the classification performance. To settle this problem, Wang *et al.*, present a new Adaptive Semi-supervised Learning model (ASL) [28] to overcome above issue, in which each unlabeled data is assigned a weight and the boundary points' weights will be suppressed for alleviating their negative effects on classification. Besides, some other manifold regularization methods also achieve awesome performance in many real applications such as Semi-supervised Orthogonal Graph Embedding (SOGE) [29], Linear Manifold Regularization with Adaptive Graph (LMRAG) [30] and so on.

However, all the aforementioned semi-supervised approaches confront a common issue that due to the restrictions of label matrix $Y \in \mathbb{R}^{n\times c}$, the reduced dimensions can only be the number of classes c, which is not suitable for high-dimensional data binary classification. In order to overcome this drawback, Cai *et al.*, [31] propose a Semi-supervised Discriminant Analysis (SDA) to maximize the separability of labeled data between different classes as well as estimate the intrinsic geometric structure through unlabeled data. However, the low-dimensional data learned by SDA is constrained in the linear subspace, which is hardly satisfied in practice. Thus, a Flexible SDA [32] based on Trace Ratio criterion (TR-FSDA) utilizes a flexible regularizer to instead the hard constraint, which can better deal with the nonlinear data. In [33], a novel Semi-supervised Orthogonal Discriminant Analysis (SODA) is designed by solving the orthogonal constrained trace ratio problem for exploring the distribution of unlabeled data effectively and learning a more discriminative subspace. A semi-supervised extension of Local Fisher Discriminant Analysis (LFDA) [21] is proposed, called SEmi-supervised Local Fisher discriminant analysis (SELF) [34] which integrates LFDA and PCA [35] into an unified model, so as to the locality of labeled data and the globality of unlabeled data can be preserved in learned subspace simultaneously. Zhang *et al.*, propose Semi-Supervised Discriminant Analysis (SSDA) [36] to maximize the separability of data between different classes via a learned robust path-based similarity measure. Nevertheless, all aforementioned methods ignore the local sub-manifold structure of data which is crucial for handling the non-Gaussian data that distributes more complex than Gaussian, that is, the samples within same class also can be divided into several sub-clusters with different manifolds.

To settle the mentioned non-Gaussian distribution problem, Zhou *et al.*, propose Semi-supervised Tangent Space Discriminant analysis (STSD) [37] which develops a new regularizer by using tangent spaces, so as to capture the local manifold structure of labeled and unlabeled data. Besides, in [10], a cluster assumption that data lie on same structure (a cluster or a sub-manifold) would have same label is proposed, thus it is key to exploit the sub-manifold in intra-class and inter-classes data simultaneously. Additionally, in [19], it claims that the data sample and its corresponding neighbor samples lie on the same sub-manifold.

Inspired by above facts, we propose a novel semi-supervised dimensionality reduction framework named Semi-supervised Adaptive Local Embedding learning (SALE), and two locality preserved semi-supervised dimensionality reduction algorithms are derived based on proposed SALE. Specifically, proposed SALE first learns a local embedding from labeled data by building k_1NN graph to explore the underlying sub-manifold structure of labeled data. Then, mapping all high-dimensional data points into learned embedding, and another k_2NN graph is constructed on all embedded data to explore integrated sub-manifolds. Intuitively, each embedded labeled data can be seen as a landmark to figure out the neighborships of unlabeled data. That is to say, each labeled point and its corresponding neighbored unlabeled points will be clustered into same sub-manifold. Last, since the data in original space contain some noises and redundant features, we learn the local embedding and k_1NN and k_2NN graphs simultaneously to overcome noisy issue.

Concretely, our contributions are fourfold:

1) We develop a novel semi-supervised adaptive local embedding dimensionality reduction framework by respectively constructing k_1NN and k_2NN graph via ℓ_0-norm constraint technique on the labeled and whole data samples for exploring local sub-manifold structure and global structure of data.
2) Two novel semi-supervised dimensionality reduction methods, named, Semi-supervised Adaptive Local Orthogonal Embedding (SALOE) and Semi-supervised Adaptive Local Whitening Embedding (SALWE) are developed based on proposed SALE framework.
3) An efficient alternatively iterative optimization algorithm is presented to solve the ℓ_0-norm constraint which is a NP-hard optimization problem.
4) Extensive experiments conducted on two synthetic data for semi-supervised dimensionality reduction and several real-world data sets for image classification achieve convincing performance, which demonstrates the effectiveness of proposed methods.

The rest of paper is organized as follows: some related works about semi-supervised learning will introduce in Section 2. In Section 3, the SALE framework is presented, then we utilize this framework to design two semi-supervised dimensionality reduction methods with orthogonal and whitening processed constraints, named SALOE and SALWE respectively. The experiments conducted on two synthetic data sets and several real-world data sets in Section 4. Finally, we conclude this paper and discuss the future works in Section 5.

2 Related Work

2.1 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a conventional dimensionality reduction method which has been employed to several real-world applications such as robust image classification [38], [39], feature selection [40] and so on.

Giving a training set $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d\times n}$, $x_i \in \mathbb{R}^d$, where n is the number of all data and d denotes the dimensionality of sample. In practice, the feature dimensionality d is very high as usual, thus the intrinsic structure of data are often hidden in low-dimensional space. As a result, how to find a transformation matrix $W \in \mathbb{R}^{d\times m}$ that projects original high-dimensional data $x_i \in \mathbb{R}^d$ to desirable low-dimensional representation $y_i \in \mathbb{R}^m$ $(m < < d)$ by $y_i = W^T x_i$ is a key point for solving the problem of "*curse of dimensionality*" [41].

The goal of classical LDA is to maximize the trace ratio value of between-class scatter matrix S_b and within-class scatter matrix S_w in the subspace [42], [43], so as to the points within same class will be pulled together while the points between different classes are pushed away as far as possible. One of LDA's objective functions is

$$\min_{W^T S_t W = I} Tr(W^T S_w W), \tag{1}$$

where $S_t = \frac{1}{n}\sum_{j=1}^{n}(x_j - \mu)(x_j - \mu)^T$ and $S_w = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n_i} (x_j^i - \mu^i)(x_j^i - \mu^i)^T$ represent the total samples scatter matrix and within-class scatter matrix respectively, and n_i denotes the number of ith class sample, x_j^i is the jth sample in ith class, μ^i denotes the mean sample of ith class and μ is the mean of all data. Note that the whitening constraint in Eq. (1), i.e., $W^T S_t W = I$, is able to enlarge the scatter of samples between different classes in the subspace. Concretely, it is well known that $Tr(S_t) = Tr(S_b) + Tr(S_w)$ which also holds for the embedded data. Moreover, $Tr(W^T S_t W)$ denotes the total samples scatter of embedded data that equals to a constant according to the constraint. Then, since Eq. (1) minimizes the scatter of embedded data within same class, i.e., $Tr(W^T S_w W)$, the scatter of embedded samples between different classes will be enlarged automatically. The above analysis is actually a whitening operation in data preprocessing, thus we call that constraint as the whitening constraint.

Without loss of generality, the within-class scatter matrix can be written as the following pairwise formulation [44], [45]

$$\widetilde{S}_w = \frac{1}{2n}\sum_{i=1}^{c}\sum_{j=1}^{n_i}\sum_{h=1}^{n_i}\frac{1}{n_i}(x_j^i - x_h^i)(x_j^i - x_h^i)^T. \tag{2}$$

Therefore, the objective function of pairwise LDA can be written to the following vector formulation:

$$\min_{W^T S_t W = I}\sum_{i=1}^{c}\sum_{j=1}^{n_i}\sum_{h=1}^{n_i}\frac{1}{n_i}||W^T x_j^i - W^T x_h^i||_2^2. \tag{3}$$

From the viewpoint of graph, there is a fully-connected graph with all weights equal to $\frac{1}{n_i}$ embedded in all pairwise points. However, it has been verified that such an objective fails to extract the local structure of data and unable to handle multimodal data actually [46].

2.2 Graph Embedding and Manifold Regularization

In the GE theory [47], an undirected graph $G \in \{V, S\}$ is constructed in which each sample is a vertex and the weights between all data pairs are stored in a similarity matrix $S \in \mathbb{R}^{n\times n}$. The similarity matrix S has various formulations, and the details can be seen in [7]. In the GE theory, an ideal case is that the smaller distance between x_i and x_j in high-dimensional space, the larger similarity between y_i and y_j in the low-dimensional subspace, vice versa. By doing so, the intrinsic neighborships of each point can be preserved via constructed graph [47]. Generally speaking, the objective of GE model can be formulated as

$$W^* = \arg\min\sum_{i\neq j}||y_i - y_j||^2 s_{ij} = \arg\min W^T X L X^T W, \tag{4}$$

where $L = D - S$ denotes Laplacian matrix of symmetrical similarity matrix S, and D is the diagonal matrix whose ith diagonal element is $D_{ii} = \sum_i s_{ij}$. Note that the similarity matrix S is obtained by the Gaussian Kernel function $exp\{-||x-y||^2/\sigma\}$ in most cases.

In real-world applications, there are a small number of labeled samples in training set, which can not satisfy the requirement of supervised methods. Without loss of generality, under the circumstance of semi-supervised, we assume the first l samples $X_l = [x_1, x_2, \ldots, x_l]$ are the labeled samples and the rest samples $\{x_{l+1}, x_{l+2}, \ldots, x_n\}$ are unlabeled. Denote the label indictor matrix of first l samples as $Y_l \in \mathbb{R}^{l\times c}$, in which the (i,j)th element of Y_l is 1 if the point x_i attributes to jth class and 0 otherwise. Ordinary Least Squares (OLS) is a classical algorithm to map high-dimensional data into c-dimensional space by using a hard approximation between regression value $(W^T X_l + \mathbf{1}b^T)$ and label matrix (Y_l). The inductive semi-supervised LR i.e., LapRLS [13] minimizes the regression errors by using labeled data and preserves the manifold smoothness on all training data simultaneously. As a result, the objective of LapRLS can be formulated as follow:

$$\min_{W,b}||W^T X_l + \mathbf{1}b^T - Y_l^T||_2^2 + \lambda_1 Tr(W^T X L X^T W) + \lambda_2||W||_2^2, \tag{5}$$

where the two parameters λ_1 and λ_2 is to balance the regression errors term and manifold regularization term, $\mathbf{1} \in \mathbb{R}^{1\times m}$ denotes a vector with all elements are 1.

Although above linear regression model has been successfully used in semi-supervised learning, the loss function is a hard approximation, i.e., the reduced dimensions can only be set as the number of classes c, which is difficult to satisfy requirements of real-world applications, such as binary classification task. The mainstream technique to address aforementioned issue is to design a new objective function based on discriminant analysis that without directly using label matrix. Next, we will review the semi-supervised discriminant analysis.

2.3 Semi-Supervised Discriminant Analysis

SDA incorporates the manifold structure by adding a regularizer term $J(a)$ to achieve the consistency assumption such that nearby original data always have similar low-dimensional representations. It is well-known that when all data samples are centralized, i.e., $\mu = \mathbf{0}$, the between-class scatter matrix equals to $S_b = X_l \widetilde{L}_b X_l^T$, where $\widetilde{L}_b$ denotes the

Laplacian matrix of $\widetilde{S}_b = diag(\widetilde{S}_b^1, \widetilde{S}_b^2, \ldots, \widetilde{S}_b^c)$ which is a symmetric matrix and denotes the graph similarity matrix where $\widetilde{S}_b^i$ $(i = 1, 2, \ldots, c)$ is a $n_i \times n_i$ matrix with all elements equal to $\frac{1}{n_i}$. Analogously, $S_w = X_l \widetilde{L}_w X_l^T$, where $\widetilde{L}_w$ denotes the Laplacian matrix of $\widetilde{S}_w$ whose (i, j)th element $\widetilde{S}_w(i, j)$ equals to $\frac{1}{l} - \widetilde{S}_b(i, j)$. The objective of SDA can be formulated as the following trace ratio problem:

$$\max_{W} \frac{Tr(W^T X_l \widetilde{L}_b X_l^T W)}{Tr(W^T (X_l(\widetilde{L}_b + \widetilde{L}_w) X_l^T + \alpha X L X^T) W)}, \tag{6}$$

where parameter α is to balance the regularizer term and L is the Laplacian matrix of whole data.

Although above semi-supervised learning methods achieve successful performance, they also suffer from two common defects: 1) The pre-defined similarity matrix is calculated in the original space, which is prone to be affected by noisy and redundant features, which fails to reveal the intrinsic structure of data. 2) The neglect of exploring sub-manifold structure leads to powerless of them on handling the data with non-Gaussian distribution.

3 Methodology

In this section, we first propose a general SALE framework, then two semi-supervised dimensionality reduction methods, named SALOE and SALWE are derived based on proposed SALE framework.

3.1 Semi-Supervised Adaptive Local Embedding Learning Framework

In this part, we introduce a novel semi-supervised adaptive local embedding learning framework based on two assumptions: 1) the intrinsic structure of high-dimensional data is hidden in the low-dimensional subspace; 2) each point and its corresponding neighbor points lie on the same sub-manifold which should be shared with same label information. Concretely, the objective of proposed SALE framework is

$$\begin{aligned} &\min_{W,P,S} f(W, P, X_l) + \alpha g(W, S, X) \\ &s.t. \quad \mathbf{1}^T \boldsymbol{p}_j^i = 1, ||\boldsymbol{p}_j^i||_0 = k_1, p_{jh}^i \geq 0, S\mathbf{1} = \mathbf{1}, \\ &\qquad\ s_{jh} \geq 0, ||\boldsymbol{s}_j||_0 = k_2, \end{aligned} \tag{7}$$

where $\boldsymbol{p}_j^i$ denotes the jth row of similarity matrix of labeled data in the ith class, and s_{jh} is the similarity value of jth sample and hth sample in all data points. In addition, the parameter α is used to control the trade-off between two terms. Specifically, $|| \cdot ||_0$ denotes the number of non-zero value in matrix, $P = diag(\boldsymbol{p}^1, \boldsymbol{p}^2, \ldots, \boldsymbol{p}^c) \in \mathbb{R}^{l \times l}$ is a similarity matrix of labeled data in which $\boldsymbol{p}^i$ denotes the similarity matrix of ith class. The ℓ_0-norm constraint, i.e., $||\boldsymbol{p}_j^i||_0 = k_1$ is to ensure that each row of matrix P only has k_1 non-zero value, which imposes the data sample x_j^i only has k_1 neighbors. Similarly, the constraint $||\boldsymbol{s}_j||_0 = k_2$ also ensures k_2 connections of sample x_j on labeled and unlabeled data for exploring integrated sub-manifold structure of all samples. Besides, the first term aims to explore the intrinsic local structure of labeled data by constructing a k_1NN graph to maintain the neighborships on each embedded sample. The second term is an unsupervised clustering term which aims to explore the global structure of all labeled and unlabeled data by constructing another k_2NN graph at learned embedded subspace. By doing so, the embedded unlabeled samples and their neighbored embedded labeled samples will be clustered into same sub-manifold as well as shared with same label information. It is worth noting that the subspace W and two graphs P and S are alternatively optimized. That is, the neighborships among all data samples are constructed on learned subspace, which can avoid the influence of noisy and redundant features on the distance calculation between each sample pairwise. In what follows, we will derive two semi-supervised dimensionality reduction methods named SALOE and SALWE based on proposed SALE framework.

3.2 Semi-Supervised Adaptive Local Orthogonal & Whitening Embedding Learning

In this section, we present two novel locality preserved semi-supervised dimensionality reduction methods based on proposed SALE framework. We first introduce a locality preserved supervised LDA model which is able to learn a local discriminative embedding by constructing k_1NN graph on labeled data. Specifically, it assigns a weight p_{jh}^i between pairwise points x_j^i and x_h^i to achieve the ideal case that the small distance in embedding space $||W^T(x_j^i - x_h^i)||_2^2$ should be assigned a large weight. Therefore, the objective of locality preserved LDA with orthogonal constraint which is beneficial to explore discriminative information [48], [49], [50], can be written as

$$\begin{aligned} &\min_{W,P} \sum_{i=1}^{c} \sum_{j=1}^{n_i} \sum_{h=1}^{n_i} (p_{jh}^i)^2 ||W^T(x_j^i - x_h^i)||_2^2 \\ &s.t. \quad \mathbf{1}^T \boldsymbol{p}_j^i = 1, p_{jh}^i \geq 0, ||\boldsymbol{p}_j^i||_0 = k_1, W^T W = I. \end{aligned} \tag{8}$$

Wherein, the reason why we use the square weights $(p_{jh}^i)^2$ rather than p_{jh}^i is to avoid the trivial solution, i.e., if the nearest neighbors of jth sample is the hth sample in the ith class, the value of p_{jh}^i approaches to 1, meanwhile the other $k_1 - 1$ nonzero elements in $\boldsymbol{p}_j^i$ approach to zero. In this circumstance, the model is prone to get stuck in local optimum, which degrades the subsequent performance.

Consequently, in order to explore the sub-manifold structure of all samples, we design to project all data samples into subspace first, and then adaptively construct another k_2NN graph S on all embedded data, such that the labeled and unlabeled data samples with neighborships will be clustered into a same sub-manifold and shared with same label information. Analogously, a smaller distance between arbitrary sample pairwise in the embedded space, i.e., $||W^T(x_j - x_h)||_2^2$ should has a larger weight s_{jh}, vice versa. To be specific, the objective function of proposed SALOE model can be formulated as

$$\begin{aligned} \min_{W,P,S} &\sum_{i=1}^{c} \sum_{j=1}^{n_i} \sum_{h=1}^{n_i} (p_{jh}^i)^2 ||W^T(x_j^i - x_h^i)||_2^2 \\ &\qquad + \alpha \sum_{j=1}^{n} \sum_{h=1}^{n} s_{jh}^2 ||W^T(x_j - x_h)||_2^2 \\ s.t. \quad &\mathbf{1}^T \boldsymbol{p}_j^i = 1, ||\boldsymbol{p}_j^i||_0 = k_1, p_{jh}^i \geq 0, S\mathbf{1} = \mathbf{1}, \\ &s_{jh} \geq 0, ||\boldsymbol{s}_j||_0 = k_2, W^T W = I. \end{aligned} \tag{9}$$

Additionally, according to the statement in the Section 2.1, the whitening constraint is used to enlarge the scatter of embedded data between different classes in the subspace, which is beneficial to improving the subsequent classification performance. Therefore, we expand our SALE model via incorporation with whitening constraint as follow:

$$\begin{aligned}\min_{W,P,S} \sum_{i=1}^{c}\sum_{j=1}^{n_i}\sum_{h=1}^{n_i} (p_{jh}^i)^2 \|W^T(x_j^i - x_h^i)\|_2^2 \\ + \alpha \sum_{j=1}^{n}\sum_{h=1}^{n} s_{jh}^2 \|W^T(x_j - x_h)\|_2^2 \quad (10) \\ s.t. \quad \mathbf{1}^T \boldsymbol{p}_j^i = 1, \|\boldsymbol{p}_j^i\|_0 = k_1, p_{jh}^i \geq 0, S\mathbf{1} = \mathbf{1}, \\ s_{jh} \geq 0, \|\boldsymbol{s}_j\|_0 = k_2, W^T S_t W = I.\end{aligned}$$

It is worth noting that, proposed SALOE and SALWE methods have three variables need to be optimized simultaneously, and the ℓ_0-norm constraint is a NP-hard problem. Thus, optimizing above models in Eqs. (9) and (10) is a challenging task. Nevertheless, we will propose an alternatively iterative optimization algorithm to solve problem (9) in the next section.

3.3 Optimization Strategy

First, we should initialize two similarity matrix P and S. In fact, the similarity matrix of labeled data $P = diag(\boldsymbol{p}^1, \boldsymbol{p}^2, \ldots, \boldsymbol{p}^c) \in \mathbb{R}^{l \times l}$ is a block diagonal matrix where $\boldsymbol{p}^i \in \mathbb{R}^{n_i \times n_i}$ $(i = 1, 2, \ldots, c)$ denotes the ith class similarity matrix. We find k_1 nearest neighbors for each point in the same class by calculating their euclidean distance in the original space, and the weights between each pairwise points can be initialized as

$$p_{jh}^i = \begin{cases} \frac{1}{k_1}, & \text{if} \quad x_h^i \in \mathcal{N}_{k_1}(x_j^i) \\ 0, & \text{otherwise}, \end{cases} \tag{11}$$

where $\mathcal{N}_{k_1}(x_j^i)$ denotes the set of k_1 nearest neighbors of x_j^i in ith class. Analogously, we initialize the elements in S as

$$s_{jh} = \begin{cases} \frac{1}{k_2}, & \text{if} \quad x_h \in \mathcal{N}_{k_2}(x_j) \\ 0, & \text{otherwise} \end{cases}, \tag{12}$$

where $\mathcal{N}_{k_2}(x_j)$ denotes the set of k_2 nearest neighbors of x_j in all samples. Next, we present the details of alternate iterative optimization algorithm for solving problem (9) in what follows.

When P and S are fixed, we solve the projections W, and the problem (9) is reduced to solve

$$\begin{aligned}\min_{W^TW=I} \sum_{i=1}^{c}\sum_{j=1}^{n_i}\sum_{h=1}^{n_i} (p_{jh}^i)^2 \|W^T(x_j^i - x_h^i)\|_2^2 \\ + \alpha \sum_{j=1}^{n}\sum_{h=1}^{n} s_{jh}^2 \|W^T(x_j - x_h)\|_2^2.\end{aligned} \tag{13}$$

According to the deductions in Section 2.2, we can rewrite problem (13) as

$$\min_{W^TW=I} Tr(W^T X_l \widetilde{L}_p X_l^T W) + \alpha Tr(W^T X \widetilde{L}_s X^T W), \tag{14}$$

where $\widetilde{L}_p = \widetilde{D}_p - \frac{\widetilde{P}^T + \widetilde{P}}{2}$ is the Laplacian matrix, $\widetilde{P} = P^2$ and the degree matrix $\widetilde{D}_p \in \mathbb{R}^{l \times l}$ is a diagonal matrix where the ith diagonal element is $\sum_h (\widetilde{p}_{jh} + \widetilde{p}_{hj})/2$. Analogously, Laplacian matrix $\widetilde{L}_s = \widetilde{D}_s - \frac{\widetilde{S}^T + \widetilde{S}}{2}$ and $\widetilde{S} = S^2$ where S denotes the similarity matrix of both labeled and unlabeled data. The Lagrange function of problem (14) is

$$\begin{aligned}\mathcal{L}(W, \lambda) = Tr[W^T(X_l \widetilde{L}_p X_l^T + \alpha X \widetilde{L}_s X^T)W] \\ - \lambda Tr(W^T W - I),\end{aligned} \tag{15}$$

where λ is the Lagrange multiplier. Setting the derivative of Eq. (15) with respect to W to zero, then we get

$$\begin{aligned}\frac{\partial \mathcal{L}(W, \Lambda)}{\partial W} = (X_l \widetilde{L}_p X_l^T + \alpha X \widetilde{L}_s X^T)W - W\Lambda = 0 \\ \Rightarrow (X_l \widetilde{L}_p X_l^T + \alpha X \widetilde{L}_s X^T)W = W\Lambda,\end{aligned} \tag{16}$$

where Λ is the Lagrange multiplier matrix. The optimal solution W to the problem (16) is formed by the m eigenvectors of $(X_l \widetilde{L}_p X_l^T + \alpha X \widetilde{L}_s X^T)$ corresponding to its m smallest eigenvalues.

When W and S are fixed, the second term of problem (9) becomes to a constant and it reduces to solve

$$\begin{aligned}\min_{P} \sum_{i=1}^{c}\sum_{j=1}^{n_i}\sum_{h=1}^{n_i} (p_{jh}^i)^2 \|W^T(x_j^i - x_h^i)\|_2^2 \\ s.t. \quad \mathbf{1}^T \boldsymbol{p}_j^i = 1, \|\boldsymbol{p}_j^i\|_0 = k_1, p_{jh}^i \geq 0.\end{aligned} \tag{17}$$

Problem (17) can be optimized in each class and it reduces to solve

$$\begin{aligned}\min_{\boldsymbol{p}_j^i} \sum_{h=1}^{n_i} (p_{jh}^i)^2 \|W^T(x_j^i - x_h^i)\|_2^2 \\ s.t. \quad \mathbf{1}^T \boldsymbol{p}_j^i = 1, \|\boldsymbol{p}_j^i\|_0 = k_1, p_{jh}^i \geq 0.\end{aligned} \tag{18}$$

It seems that the ℓ_0-norm constraint is difficult to be satisfied, which is a NP-hard optimization problem. ℓ_0-norm constraint optimization algorithm as the major theoretical contribution of this paper will be introduced in the following contents. Without loss of generality, we assume a column vector $\boldsymbol{\eta}$ to replace $\boldsymbol{p}_j^i$ and v_h denotes $\|W^T(x_j^i - x_h^i)\|_2^2$. Then the problem (18) becomes to the following simple formulation

$$\min_{\mathbf{1}^T\boldsymbol{\eta}=1, 0 \leq \eta, \|\boldsymbol{\eta}\|_0 = k_1} \sum_{t=1}^{n_i} \eta_t^2 v_t. \tag{19}$$

The constraint $\|\boldsymbol{\eta}\|_0 = k_1$ will impose that there have only k_1 nonzero values in vector $\boldsymbol{\eta}$. Therefore, we assume the indexes of those k_1 nonzero values in $\boldsymbol{\eta}$ are $\{r_1, r_2, \ldots, r_t, \ldots, r_{k_1}\}$ $(1 \leq r_t \leq n_i, r_t \neq j)$. Then the above problem reduces to

$$\min_{\mathbf{1}^T\boldsymbol{\eta}=1, 0 \leq \eta} \sum_{t=1}^{k_1} \eta_{r_t}^2 v_{r_t}, \tag{20}$$

in which the ℓ_0-norm constraint has been subtly circumvented. Without the constraint $0 \leq \eta$, the Lagrange function of problem (20) is

$$\mathcal{L}(\eta, \lambda) = \sum_{t=1}^{k_1} \eta_{r_t}^2 v_{r_t} - \lambda\Big(\sum_{t=1}^{k_1} \eta_{r_t} - 1\Big). \tag{21}$$

Setting the derivative of Eq. (21) in terms of η_{r_t} to zero and combining the constraint $\mathbf{1}^T\boldsymbol{\eta} = 1$, then we can obtain the following closed-form solution of problem (20)

$$\eta_{r_t} = \frac{1}{v_{r_t}} \times \Big(\sum_{t=1}^{k_1} \frac{1}{v_{r_t}} \Big)^{-1}. \tag{22}$$

Note that, since the Eq. (22) always satisfy the $0 \leq \eta$ constraint, the Eq. (22) is an optimal solution of problem (20). Substituting Eq. (22) into Eq. (19), the objective function value of Eq. (19) is

$$\sum_{h=1}^{n_i} \eta_h^2 v_h = \frac{1}{\big(\sum_{t=1}^{k_1} v_{r_t}^{-1}\big)^2} \sum_{t=1}^{k_1} v_{r_t}^{-1} = \frac{1}{\big(\sum_{t=1}^{k_1} v_{r_t}^{-1}\big)}. \tag{23}$$

Notice that the monotonicity of function value of Eq. (23) is consistent with variable v_{r_t}. Thus, in order to minimize the Eq. (19), the v_{r_t} $(t = 1, 2, \ldots, k_1)$ should be the k_1 smallest values in the vector $\mathbf{v}$, i.e., the indexes of k_1 nearest neighbors of x_j^i in the subspace. Denoting r_t^* $(t = 1, 2, \ldots, k_1)$ as the indexes of k_1 smallest values in the vector $\mathbf{v}$ and the optimal solution of problem (18) is

$$p_{jh}^i = \begin{cases} \frac{(||W^T(x_j^i - x_{r_t^*}^i)||_2^2)^{-1}}{\sum_{t=1}^{k_1}(||W^T(x_j^i - x_{r_t^*}^i)||_2^2)^{-1}}, & \text{if} \quad h = r_t^* \\ 0, & \text{otherwise.} \end{cases} \tag{24}$$

It is worth noting that the weight p_{jh}^i between intra-class points x_j^i and x_h^i is not calculated in the original space, but in the learned subspace. That is to say, with the update of subspace W, the weights will become more accurate. Besides, it is evident that the smaller the distance $||W^T(x_j^i - x_{r_t}^i)||_2^2$ is, the larger the weight p_{jh}^i becomes as well, which is consistent with our previous assumption and verifies the effectiveness of our solution in Eq. (24).

When W and P are fixed, the first term of problem (9) becomes a constant, and it reduces to solve

$$\begin{aligned} \min_{S} & \sum_{j=1}^{n} \sum_{h=1}^{n} s_{jh}^2 ||W^T(x_j - x_h)||_2^2 \\ s.t. \quad & S\mathbf{1} = \mathbf{1}, s_{jh} \geq 0, ||\boldsymbol{s}_j||_0 = k_2. \end{aligned} \tag{25}$$

For each point x_j, the above problem (25) equals to solve

$$\begin{aligned} \min_{\boldsymbol{s}_j} & \sum_{h=1}^{n} s_{jh}^2 ||W^T(x_j - x_h)||_2^2 \\ s.t. \quad & \boldsymbol{s}_j^T \mathbf{1} = 1, s_{jh} \geq 0, ||\boldsymbol{s}_j||_0 = k_2. \end{aligned} \tag{26}$$

It is evident that solving problem (26) is similar to solving the problem in Eq. (17). Therefore, we can provide the optimal solution of problem (26) directly as follow:

$$s_{jh} = \begin{cases} \frac{(||W^T(x_j - x_{q_t^*})||_2^2)^{-1}}{\sum_{t=1}^{k_2}(||W^T(x_j - x_{q_t^*})||_2^2)^{-1}}, & \text{if} \quad h = q_t^* \\ 0, & \text{otherwise,} \end{cases} \tag{27}$$

where q_t^* represents the indexes of k_2 nearest neighbors of point x_j in labeled and unlabeled data.

Besides, in the optimization of SALWE model, the solutions of P and S are same to the ones in the SALOE model, but the columns of projections W are composed by m eigenvectors of $S_t{}^{-1}(X_l \widetilde{L}_p X_l^T + \alpha X \widetilde{L}_s X^T)$ corresponding to m smallest eigenvalues. In practice, we can leverage PCA preprocessing to remove the null space of S_t, so as to guarantee the reversibility of S_t matrix.

So far, the optimization steps of three variables have been described at all. we summarize our optimization algorithm in the following Algorithm 1. It is easy to prove that the following Algorithm 1 will decrease the objective function value in Eq. (9) in each iteration. Besides, proposed objective function in Eq. (9) has a lower bound 0, and the convergent condition used in our experiments is set as $|obj^t - obj^{t+1}| \leq 10^{-4}$.

Algorithm 1. Optimization Procedure of Problem (9)

```
Input : $X = [x_1, x_2, \ldots, x_l, x_{l+1}, \ldots, x_n] \in \mathbb{R}^{d \times n}$,
        $X_l = [x_1, x_2, \ldots, x_l] \in \mathbb{R}^{d \times l}$, $\boldsymbol{y} \in \mathbb{R}^l$, $k_1$, $k_2$, $m$.
Output : Transformation matrix: $W \in \mathbb{R}^{d \times m}$.
Initialisation: $P \in \mathbb{R}^{l \times l}$ and $S \in \mathbb{R}^{n \times n}$;
while not convergence do
    $\widetilde{L}_p \leftarrow P, \quad \widetilde{L}_s \leftarrow S$ ;
    $W$[1] $\leftarrow \mathsf{eigs}((\mathsf{X_l \widetilde{L}_p X_l^T} + \alpha \mathsf{X \widetilde{L}_s X^T}), '\mathsf{descend}', \mathsf{m})$;
    Update $P$ according to Eq. (24);
    Update $S$ according to Eq. (27);
end
```

4 Experiments

In this section, we first provide two visualization experiments to verify the superiority of our proposed method on discriminative subspace learning. Then, we evaluate the performance of proposed SALWE and SALOE algorithms on real-world data sets classification tasks by comparison with several related SOTA semi-supervised dimension reduction methods. Last, we provide two studies in terms of convergence and parameters sensitivity of our methods.

4.1 Visualization Experiments

In this section, we conduct visualization experiment for evaluating the ability of dealing with non-Gaussian data and exploring the local data structure of proposed models on two synthetic data sets. After that, an visualization of learned graph is provided.

We first conduct an visualization experiment on synthetic 2-dimensional non-Gaussian data for evaluating the ability of exploring local structure. In Fig. 1, the red circles and blue squares denote the samples from two different classes. Wherein, the samples of class 1 that divided into two subclusters does not obey the Gaussian distribution, which is the so-called non-Gaussian data. We perform four semi-supervised dimensionality reduction algorithms, i.e., SDA, TR-FSDA and proposed SALWE and SALOE models on such a 2-dimensional non-Gaussian data set for learning the 1-dimensional projection. From the results, we can observe that two global algorithm, SDA and TR-FSDA fail to find an appropriate projection, so as to the embedded samples will overlapping each other. In contrast, since proposed SALWE and SALOE algorithms are able to explore the local structure from the labeled non-Gaussian data by constructing k_1NN graph, the samples mapped into the learned projection are

[1] For satisfying the constraint $W^T W = I$, updating W as $W = W \times diag(1./\sqrt{diag(W^T W)})$.

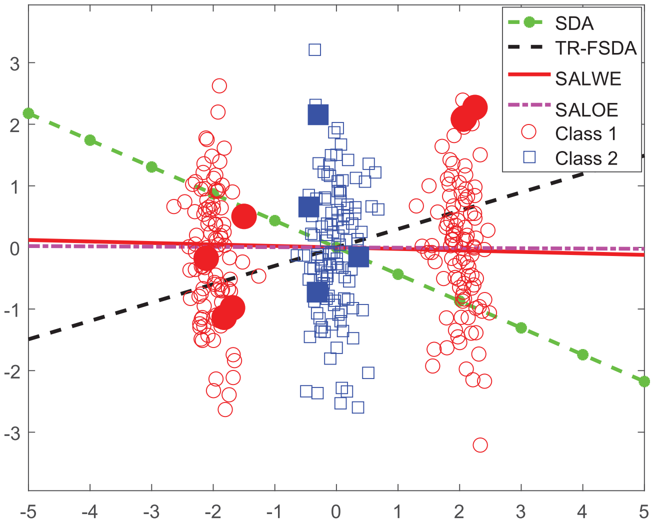
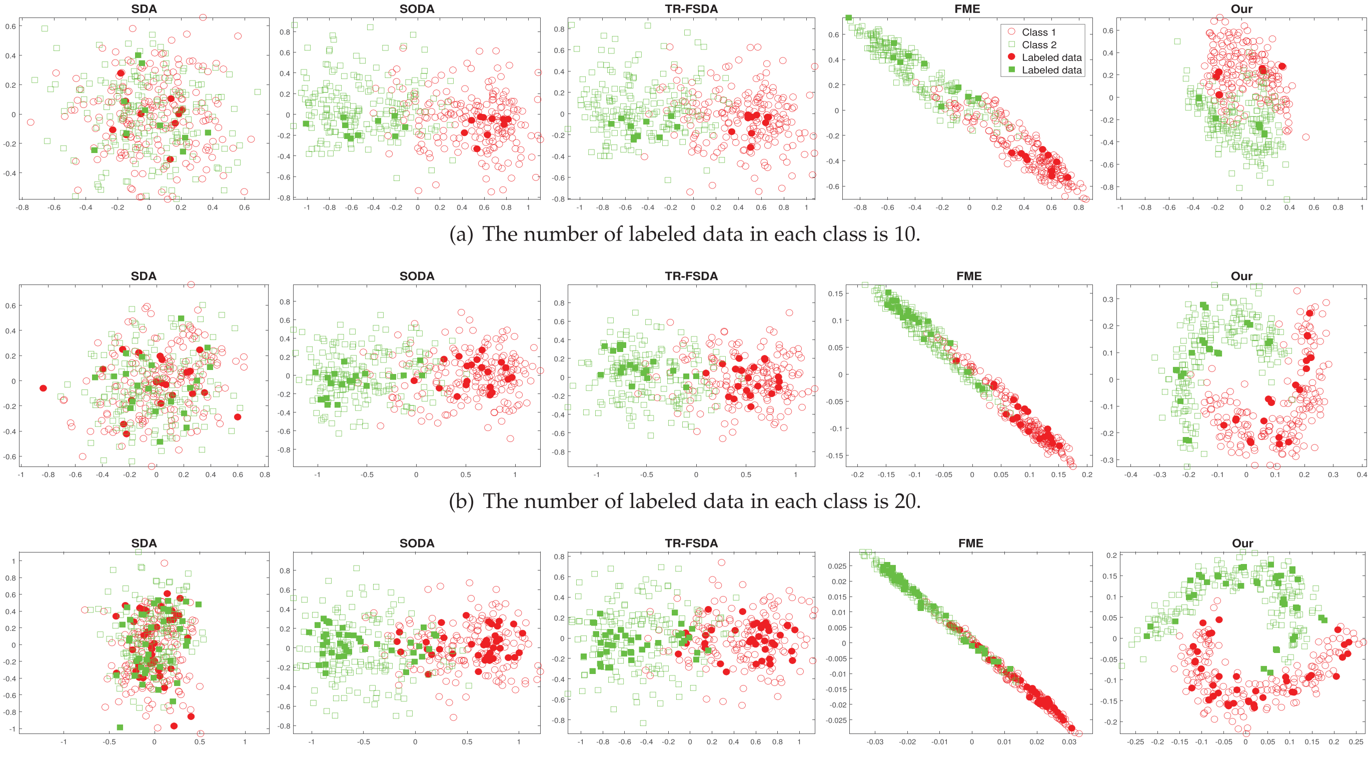


Fig. 1. The synthetic 2-dimensional non-Gaussian data example. Four projections are generated by SDA (green dashed line), TR-FSDA (black dashed line), SALWE (red line) and SALOE (pink dashed line), respectively.

well-separated. This experiment demonstrates the effectiveness of our methods on handling non-Gaussian data.

Additionally, we conduct another visualization experiment on the two-moon synthetic data which contains 400 samples with two non-Gaussian distributed classes (red circle and green square). In order to simulate the circumstance of dimensionality reduction, we add eight-dimensional Gaussian noises on all two moon data. Then SALOE and four compared semi-supervised dimensionality reduction algorithms, including SDA, SODA, TR-FSDA and FME are applied to project the generated 10-dimensional data samples into 2-dimensional subspace. Moreover, we conduct two moon data experiment with different number of labeled data (solid points), and the embedded results are shown in Fig. 2. The key for judging the experimental results is whether the manifold structure of unlabeled data points, i.e., the two-moon shape can be entirely explored. From the Fig. 2, we can observe that SDA fails to explore the two-moon local structure no matter how many labeled samples there are. Moreover, although there are partially overlap occurred in the subspace produced by SODA and TR-FSDA, the two-moon shape are always missing. In Fig. 2a, as the labeled points are inadequate, the performance of our method is similar to SODA and TR-FSDA, instead, FME works well in which the distribution of embedding points look like a circle, but it still far away from two moon shape. Furthermore, in Figs. 2b and 2c, with the increase of labeled data point, the distribution of 2-dimensional data generated by our method is close to two-moon shape, but FME still does not work well. Experimental results verify the effectiveness of our methods on exploring local and global data structure by comparison with related semi-supervised learning methods.

Last, in Fig. 3, we present an visualization of k_2NN ($k_2 = 3$) graph learned by SALWE and SALOE on MSRA25, PIE and YaleB data sets respectively. For intuition, we use the first ten classes of each data and each class contains 10 samples as input of algorithms. Moreover, the data points are reorganized such that the samples with same label are placed together. The prototype of kNN graph is obtained by Gaussian kernel function, and the similarity of our methods is calculated according to Eq. (27). It is obvious that kNN graph has slight block diagonal structure while the two optimal graphs have excellent structural characters on MSRA25 and PIE data sets, which demonstrates that the neighbors selected by Eq. (27) are more accurate than Gaussian kernel function in pre-defined graph. Besides, the kNN graph on YaleB data set

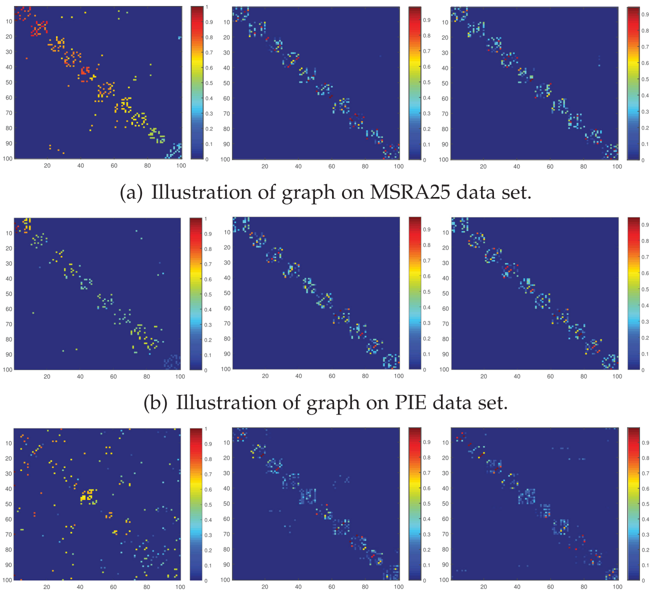


(a) The number of labeled data in each class is 10.

(b) The number of labeled data in each class is 20.

(c) The number of labeled data in each class is 40.

Fig. 2. Visualization on 2D embeddings of two moon data with different number of labeled data.



(a) Illustration of graph on MSRA25 data set.

(b) Illustration of graph on PIE data set.

(c) Illustration of graph on YaleB data set.

Fig. 3. Illustration of the kNN graph (left column) and two optimal graph learned by SALWE (middle column) and SALOE (right column) on MSRA25 (a), PIE (b), YaleB (c) respectively.

is disorder while our optimal graphs have clear block diagonal structure, although there are few inaccurate points as well. The above experimental results verify the success of our framework on adaptive graph construction strategy.

4.2 Real-World Data Sets Classification

In this section, classification experiments are conducted on several real-world data sets with comparing to nine SOTA semi-supervised learning algorithms to evaluate the ability of extracting discriminative structure.

4.2.1 Data Sets

We use five real-world data sets to conduct classification experiments. More detailed descriptions will be presented in what follows.

- *Coil20* [51] data set, named Columbia Object Image Library contain 16×16 grayscale images of 20 objects which have a wide variety of complex geometric, appearance and reflectance characteristics.
- *MSRA25* [52] face data set collected by the Microsoft Research Asia that contain 12 individuals, and each class includes 113-186 images.
- *AR* [53] face data set contains over 4,000 color images corresponding to 126 people's faces (70 men and 56 women). Images feature frontal view faces with different facial expressions, illumination conditions and occlusions. We use a subset (50 men and 50 women, each person has 26 images) of AR and the color images are converted into 165×120 gray images.
- *YaleB* [54] data set is an extension of Yale data set. In our experiment, we simply use the cropped images which contain 32×32 pixels. This data set now has 38 individuals and around 64 near frontal images under different illuminations per individual.
- *CMU-PIE* [55] data set contains 41,368 images of 68 people, in which each person is presented in 13 different poses, under 43 different illumination conditions, and with 4 different expressions. In our experiment, we only use one pose from PIE, named POSE C29 which consists of 1,632 samples with 68 classes.

4.2.2 Compared Algorithms and Parameters Settings

To verify the superiority of proposed method in the aspect of extracting intrinsic structure, we compare the following nine state-of-the-art semi-supervised learning algorithms:

- *SDA* [31] extends LDA to semi-supervised version by adding a regularizer term, and the regularization parameter α is set from 10^{-3} to 10^{3} in our experiments.
- *TR-FSDA* [32] employs a flexible regularizer term that models the regression residual into the objective function, which is able to explore the non-linear structure of data. A regularization parameter is tuned from 10^{-3} to 10^{3} as well.
- *SODA* [33] improves the discriminability of subspace by solving an orthogonal constrained trace ratio problem, and we tune the regularization parameter similar to above method.
- *SELF* [34] has two alternate parameters need to be set, wherein the type of metric in the embedding space is set to "Weighted" and the affinity parameter used in local scaling heuristic is set to default value 7.
- *STSD* [37] suggests the parameter of regularizer should be searched from 0 to 1 with 0.1 interval, and other parameters including the parameter within tangent based regularizer γ, the trade off parameter for the Tikhonov regularizer β are set to default value 1 and 0.0001 respectively.
- *SSDA* [36] has three regularization parameters, namely α, β and γ which are tuned from 1 to 1.1 with 0.1 interval, set to 10^{-3} and fixed to 0.1 respectively.
- *FME* [26] that is a framework for relaxing the hard constraint in traditional LapRLS algorithm. However, it has four parameters need to be tuned and we set them in range of $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$.
- *ASL* [28] addresses the boundary points issue by assigning small weights to the unlabeled data which lie on the boundary line. According to the suggestion in ASL, we select the parameter r from 1 to 2 with 0.1 interval.
- *SOGE* [29] integrates the manifold smoothness and a penalization term to learn an orthogonal projection and recursively update the projection matrix for improving the dimensionality reduction performance. For fair comparison, we tune the regularization parameter in the set of $\{10^{-9}, 10^{-6}, 10^{-3}, 10^{0}, 10^{3}, 10^{6}, 10^{9}\}$.

Besides, for proposed methods, we tune the regularization parameter α using the "grid-search" strategy from the range $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$. For simplicity, the number of neighbors in within-class k_1 and between-classes k_2 are searched from 1 to $(n_i - 1)$ with 1 interval, where n_i is the minimal number of samples among all classes and fixed to 10, respectively.

4.2.3 Preliminary

In order to remove the null space of original data, we adopt PCA preprocessing on these data to preserve 95 percent

TABLE 1
Recognition Performance (Mean Recognition Accuracy ± Standard Deviation %) of SDA, TR-FSDA, SODA, SELF, STSD, SSDA, FME, ASL, SOGE, SALWE and SALOE on Five Data Sets

Data	Method	10% Labeled		20% Labeled		30% Labeled		40% Labeled		50% Labeled	
		Unlabel	Test	Unlabel	Test	Unlabel	Test	Unlabel	Test	Unlabel	Test
Coil20	SDA	77.66±2.40	78.35±1.87	87.52±2.59	87.82±2.07	93.60±0.77	94.25±1.05	95.30±1.77	95.93±1.02	97.86±0.45	98.08±0.32
	TR-FSDA	84.44±2.16	84.40±2.16	92.05±1.93	91.72±2.40	95.24±0.76	95.00±1.19	96.61±1.29	96.56±0.63	98.53±0.62	98.18±0.44
	SODA	**84.55**±2.00	**84.65**±2.03	**92.07**±1.93	**91.75**±2.74	95.20±1.08	**95.08**±1.20	96.68±1.19	96.48±0.77	98.47±0.53	98.02±0.57
	SELF	80.16±1.56	79.00±2.42	85.52±2.37	86.44±2.14	90.04±1.69	90.72±1.10	92.58±0.87	92.73±1.35	93.72±1.58	94.06±0.46
	STSD	84.38±1.41	83.64±1.62	89.59±1.71	90.33±1.36	94.04±1.41	94.97±1.74	96.68±1.30	**97.06**±0.82	97.67±0.93	97.50±0.78
	SSDA	81.47±3.56	81.81±2.21	90.24±1.53	90.50±1.12	94.60±0.53	94.67±1.24	96.70±0.76	96.75±0.68	98.22±1.28	98.14±0.45
	FME	79.59±1.83	82.19±1.11	83.21±2.38	82.26±2.63	86.00±1.06	86.19±1.23	86.77±2.34	87.18±2.54	88.72±1.82	88.86±1.57
	ASL	77.97±2.29	75.01±0.90	82.71±2.92	81.75±2.63	86.92±1.47	86.79±1.22	88.05±2.18	88.54±1.65	90.19±1.32	90.58±1.31
	SOGE	78.95±2.09	79.24±2.40	85.90±2.10	86.57±1.98	91.34±0.76	91.58±0.65	93.64±1.21	93.85±1.11	96.39±1.11	96.88±0.70
	SALWE	74.61±2.10	74.92±2.49	86.84±2.69	87.29±2.59	94.06±1.17	94.06±1.76	**96.73**±0.81	96.68±0.47	**98.86**±0.29	**98.24**±0.28
	SALOE	83.05±2.28	83.47±2.01	90.17±1.90	90.03±1.98	**95.88**±0.96	94.94±1.51	96.55±1.34	96.75±0.70	98.33±0.62	98.21±0.44

Data	Method	10% Labeled		20% Labeled		30% Labeled		40% Labeled		50% Labeled	
		Unlabel	Test	Unlabel	Test	Unlabel	Test	Unlabel	Test	Unlabel	Test
MSRA25	SDA	80.47±3.24	80.95±1.96	87.26±1.60	87.56±1.12	89.46±1.39	88.62±1.07	90.74±1.24	90.70±0.96	91.16±1.39	91.80±0.81
	TR-FSDA	80.71±2.99	80.76±1.76	87.04±1.49	87.29±1.28	89.42±1.37	88.79±0.81	90.65±1.11	90.54±0.90	91.00±1.25	91.43±0.64
	SODA	81.07±2.92	80.95±1.87	87.23±1.45	87.49±1.49	89.42±1.02	88.81±0.91	90.87±1.29	90.47±0.94	90.87±1.34	91.43±0.71
	SELF	77.98±1.86	79.17±1.24	86.65±1.33	85.85±1.50	89.61±1.04	88.39±1.44	89.72±0.83	90.27±1.00	91.76±0.93	91.34±0.34
	STSD	**85.69**±1.70	**86.70**±1.79	88.03±1.65	87.75±1.01	90.24±1.43	89.46±1.30	89.87±0.78	90.36±0.67	91.22±1.33	90.87±0.41
	SSDA	77.93±2.71	77.97±3.04	85.71±1.49	85.85±1.78	88.82±0.66	88.55±1.15	90.72±1.49	89.11±0.55	90.07±1.87	89.67±1.34
	FME	81.13±4.16	69.54±2.58	85.91±2.22	74.00±3.28	87.42±1.29	74.69±2.41	87.43±1.29	76.81±1.77	87.57±1.42	78.44±2.46
	ASL	68.98±4.62	98.79±3.75	79.92±5.21	80.90±4.77	84.31±2.17	83.50±1.48	85.38±0.70	85.57±1.27	87.15±1.08	86.88±1.40
	SOGE	82.56±3.02	81.61±3.14	80.82±2.52	80.93±2.81	85.70±1.47	84.22±1.15	85.93±1.17	85.60±1.72	87.15±1.39	87.66±1.08
	SALWE	76.50±4.39	77.05±3.59	88.56±1.20	88.73±0.88	91.09±0.94	89.79±1.08	**92.14**±0.94	91.29±1.14	**92.32**±0.76	**92.60**±0.80
	SALOE	83.68±3.62	82.90±3.92	**89.81**±1.26	**89.96**±1.00	**91.20**±0.78	**90.29**±1.10	92.07±0.97	**91.90**±0.99	92.07±0.35	92.18±1.03

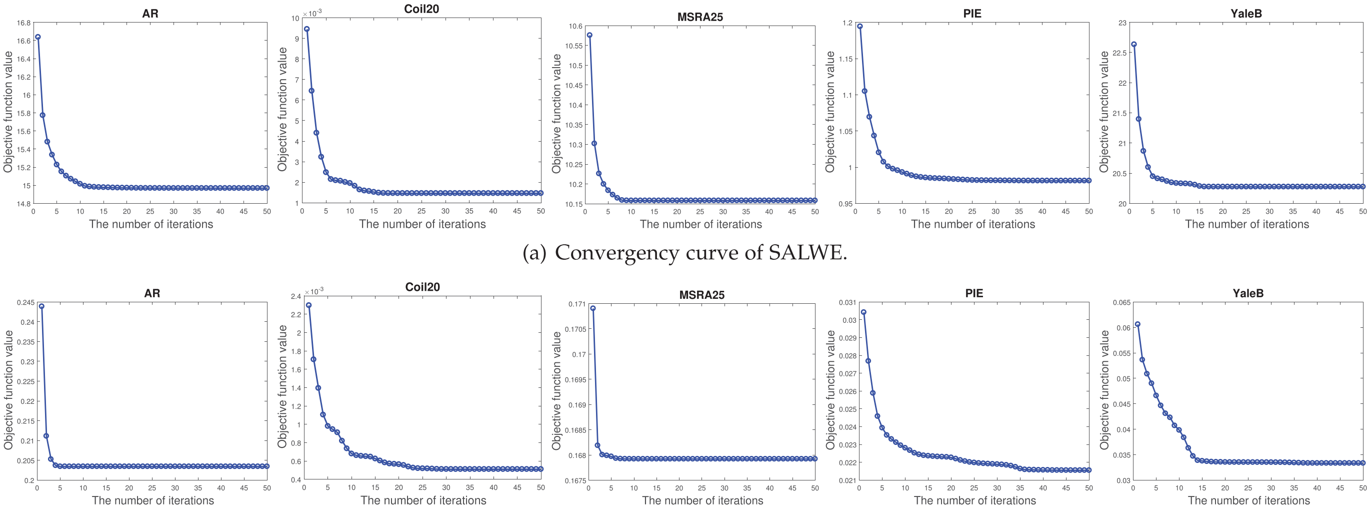
Data	Method	20% Labeled		30% Labeled		40% Labeled		50% Labeled		60% Labeled	
		Unlabel	Test	Unlabel	Test	Unlabel	Test	Unlabel	Test	Unlabel	Test
AR	SDA	59.35±1.50	61.46±1.37	71.01±2.12	72.82±1.63	77.79±1.77	79.05±1.43	86.40±2.08	86.35±1.56	90.64±1.16	**91.07**±0.89
	TR-FSDA	68.66±2.40	68.55±1.40	77.63±1.31	78.00±0.96	**82.66**±1.29	82.42±1.16	87.17±1.51	87.39±1.32	90.06±0.74	89.97±0.92
	SODA	61.01±2.62	61.14±2.05	70.13±1.85	70.03±1.18	75.17±1.05	75.61±1.92	81.62±2.28	81.87±1.25	84.58±1.08	84.45±1.54
	SELF	53.56±1.06	52.74±1.39	60.69±2.34	61.12±0.81	65.22±1.42	65.11±1.64	71.53±0.94	71.72±1.14	74.48±2.23	75.22±1.12
	STSD	51.10±0.73	51.58±1.25	63.44±2.07	63.97±1.52	69.33±2.13	70.35±1.18	77.83±0.62	77.98±0.93	80.32±2.24	82.34±1.68
	SSDA	56.84±1.02	56.75±0.70	66.18±1.07	65.52±1.18	72.17±1.74	71.88±1.72	80.90±0.79	81.80±2.15	83.48±0.88	83.65±1.46
	FME	56.29±2.47	56.32±1.32	63.04±1.66	63.37±1.23	67.80±1.88	68.01±1.52	74.33±1.15	74.04±1.29	75.70±1.26	75.46±1.23
	ASL	**71.90**±1.90	**72.19**±1.54	77.23±1.21	77.58±1.29	81.49±1.34	81.42±1.14	85.95±1.49	86.30±1.04	86.86±1.21	87.56±0.76
	SOGE	21.06±1.64	21.22±0.85	25.52±1.22	24.82±1.30	28.09±1.09	28.56±1.00	33.68±1.85	34.42±0.94	37.94±2.09	36.28±1.17
	SALWE	67.50±1.70	68.05±2.50	76.84±1.84	76.82±0.95	81.80±1.52	82.23±1.53	**88.25**±1.59	**88.12**±1.00	**90.74**±0.66	90.39±1.38
	SALOE	71.22±2.75	71.86±2.45	**78.73**±1.72	**79.68**±0.95	82.64±1.09	**82.53**±0.87	87.47±1.59	87.37±1.07	90.30±0.96	89.98±0.97

Data	Method	10% Labeled		20% Labeled		30% Labeled		40% Labeled		50% Labeled	
		Unlabel	Test	Unlabel	Test	Unlabel	Test	Unlabel	Test	Unlabel	Test
YaleB	SDA	60.78±3.04	61.34±2.41	78.53±1.98	78.83±2.14	87.47±1.35	87.55±0.93	89.99±1.14	90.42±0.51	91.59±0.74	91.75±0.58
	TR-FSDA	61.11±2.09	60.96±1.72	80.84±1.47	80.53±1.58	89.59±0.82	88.91±0.81	91.61±1.21	92.12±0.60	93.10±1.12	93.24±0.59
	SODA	49.74±2.21	49.71±1.84	72.11±1.20	71.52±2.10	82.62±1.27	81.73±1.10	86.20±1.32	86.71±0.71	88.74±1.33	88.61±0.82
	SELF	58.05±2.73	58.01±1.60	77.37±1.64	78.01±1.32	84.73±1.19	85.26±0.79	87.65±1.76	88.22±0.63	89.64±0.77	89.68±0.71
	STSD	49.04±3.95	51.00±3.71	73.13±1.85	73.53±1.28	84.92±1.91	84.48±1.69	88.98±1.62	89.51±0.85	90.79±1.27	91.19±0.63
	SSDA	58.81±3.58	57.78±1.79	76.76±0.50	77.05±0.84	87.78±0.76	86.90±0.70	90.43±2.14	90.41±1.16	92.28±1.33	92.23±1.05
	FME	52.83±1.83	52.71±1.29	64.51±2.40	63.91±1.81	73.37±1.92	72.07±1.41	76.84±2.15	76.77±1.62	80.10±1.83	79.24±1.34
	ASL	61.95±2.08	61.26±2.72	73.21±1.90	73.06±1.27	81.14±1.66	79.65±1.30	83.42±1.51	83.68±1.28	85.88±1.22	84.95±1.12
	SOGE	46.68±2.40	48.92±2.87	62.98±2.50	64.67±2.85	74.25±1.91	76.21±0.58	78.07±0.89	81.20±1.40	81.77±1.07	84.29±1.27
	SALWE	62.64±2.91	61.25±2.09	80.27±1.62	80.22±2.08	88.21±0.96	87.19±1.18	91.02±1.19	90.91±0.75	**93.81**±1.00	**93.66**±0.71
	SALOE	**66.01**±2.84	**65.36**±2.17	**82.21**±0.85	**82.11**±1.49	**90.10**±0.75	**89.24**±0.60	**91.95**±1.06	**92.15**±0.57	93.59±0.75	93.40±0.52

Data	Method	20% Labeled		30% Labeled		40% Labeled		50% Labeled		60% Labeled	
		Unlabel	Test	Unlabel	Test	Unlabel	Test	Unlabel	Test	Unlabel	Test
PIE	SDA	73.29±2.59	74.99±2.67	87.76±1.26	87.51±1.06	89.29±0.92	89.77±0.97	91.78±0.95	91.31±0.80	92.65±1.42	92.39±1.13
	TR-FSDA	75.75±2.23	76.83±2.85	90.00±0.95	**89.77**±0.60	91.28±1.26	90.98±1.15	92.77±0.96	92.46±0.82	**93.88**±1.59	**93.16**±0.88
	SODA	68.22±1.94	68.85±3.36	85.22±1.57	84.61±1.05	87.25±2.23	87.35±1.39	90.49±1.14	89.90±1.10	91.26±1.79	91.15±1.29
	SELF	77.21±1.95	76.99±2.44	88.24±1.52	87.94±1.78	89.33±1.41	89.46±0.96	90.10±1.92	91.47±0.58	91.59±0.61	91.40±0.30
	STSD	73.56±3.95	73.92±2.67	88.71±2.36	88.65±1.28	90.00±2.05	90.20±0.86	91.32±1.73	92.13±1.18	92.53±1.13	92.50±0.42
	SSDA	71.91±1.88	71.76±3.09	86.99±1.78	88.48±1.39	89.54±1.53	90.76±1.12	92.11±1.53	91.62±1.48	93.12±2.21	91.89±0.88
	FME	69.31±1.58	69.63±2.62	79.91±1.54	79.66±1.67	81.68±2.28	82.68±1.10	84.49±1.44	84.49±1.42	85.97±1.36	85.22±1.76
	ASL	78.78±1.78	78.95±2.36	88.11±1.22	88.08±0.71	88.42±1.23	89.34±0.98	90.71±0.88	90.05±0.94	91.47±1.66	91.02±1.16
	SOGE	68.43±2.97	69.96±2.85	84.56±1.51	84.44±1.16	86.39±1.50	87.07±1.12	89.56±1.53	89.09±1.14	90.76±1.72	90.36±1.00
	SALWE	77.01±2.24	77.62±2.62	**90.31**±1.06	88.57±0.88	**91.58**±1.29	**91.63**±0.90	92.85±0.69	**92.70**±0.53	93.26±1.66	92.48±1.08
	SALOE	**80.34**±2.11	**80.69**±2.39	89.89±1.06	89.46±0.81	90.84±1.44	90.75±0.96	**93.41**±0.57	92.54±0.74	93.85±1.53	93.09±1.00

The results shown in boldface are the best results in each group comparison.

information of all data. The nearest neighbor classifier is used to classify the data after dimension reduction. We repeatedly conduct all experiments with 10 times and the average classification accuracy and standard deviations under best dimensions and parameters are recorded on Table 1. Besides, in each experiment, we randomly selected 50 percent of all



(a) Convergency curve of SALWE.

(b) Convergency curve of SALOE.

Fig. 4. Convergency curves of two proposed methods over all of used data sets. (a) SALWE. (b) SALOE.

samples as training data and remaining samples for testing. Among the training data, we randomly label different proportions samples per class and treat other samples as the unlabeled data. According to the number of data in each class of different data, we label 10-50 percent data per class in Coil20, MSRA25 and YaleB data sets and label 20-60 percent data per class in AR and PIE data sets respectively.

4.2.4 *Experimental Results Analysis*

Table 1 shows the classification performance, and some observations can be summarized as follows:

- SDA, SODA and TR-FSDA have better performance than other comparing algorithms, especially, when the number of labeled data increases, TR-FSDA almost achieves the best results on AR, YaleB and PIE data sets, SODA works well on Coil20 data set, and SDA is generally better on MSRA25 data set, in terms of mean classification accuracy.
- Proposed methods do not achieve the best recognition accuracy on the Coil20 data set when the number of labeled data is small, because a small amount of labeled data hardly represent the intrinsic submanifold of whole points, so as to the local discriminative embedding can not be explored accurately.
- Proposed methods outperform several global methods including SDA, SODA and FME in most cases, which demonstrates the effectiveness of locality preserved property of proposed framework. In other words, our proposed semi-supervised dimensionality reduction has more power on exploring the discriminative information from high-dimensional data. Additionally, proposed methods have better performance than SELF, SSDA and STSD as well, which demonstrates the superiority of proposed methods in terms of exploring local structure of data samples.

4.3 Convergence Study

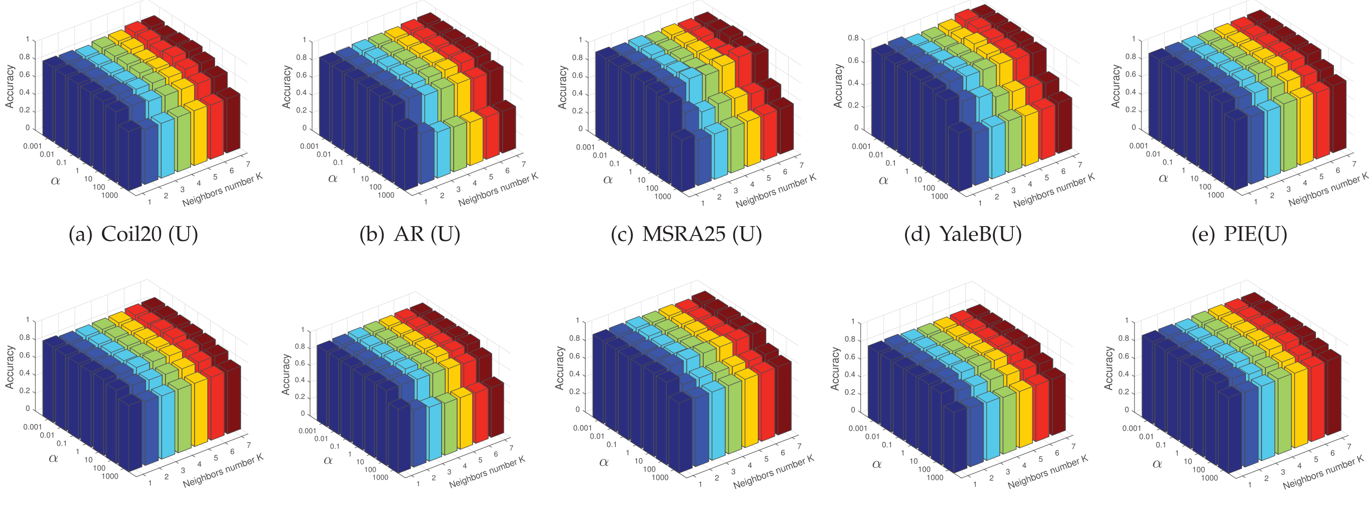
In the section, we provide the convergence curves of two proposed methods over all used data sets, the regularization parameter is set to 0.1 and the number of neighbors in labeled (k_1) and whole data (k_2) are fixed at 2 and 10 respectively in each experiment. From the Fig. 4, it is obvious that our iterative optimization algorithm converges to the optimum within less than 20 iterations on most of data sets. Moreover, on the AR and MSRA25 data sets, proposed methods converges only within less than 10 iterations. That is, two proposed algorithms can efficiently achieve convergence in practice.

4.4 Sensitivity Study

In this section, we evaluate the sensitivity of parameters α, k_1 and initialization P, S in our proposed models. First, we study the regularization parameter α and the number of neighbors k_1 in SALWE method for classifying unlabeled data (first row in Fig. 5) and test data (second row in Fig. 5). We randomly select 50 percent of all samples for training and rest of samples for testing, then choose 20 percent samples of training set as labeled data over all data sets. From the Fig. 5, we can observe that the performance of proposed method show stability and efficiency within the range of $\alpha = \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10,\ 10^{2}, 10^{3}\}$, especially when the value of k_1 is small in most case. Therefor, the performance of our SALWE model is robust to the regularization parameter α, but we still suggest to perform hierarchy grid search strategy to get better result in real-world applications.

Second, we conduct another sensitivity analysis experiment on proposed SALOE model for studying the relationship between the regularization parameter α and the number of labeled data in training set. From the Fig. 6, we can conclude that our method always achieve better performance when the value of α is small, which is in line with the conclusion in Fig. 5. Besides, the tendency of performance with the change of parameter α is not affected by the varying number of labeled data actually.

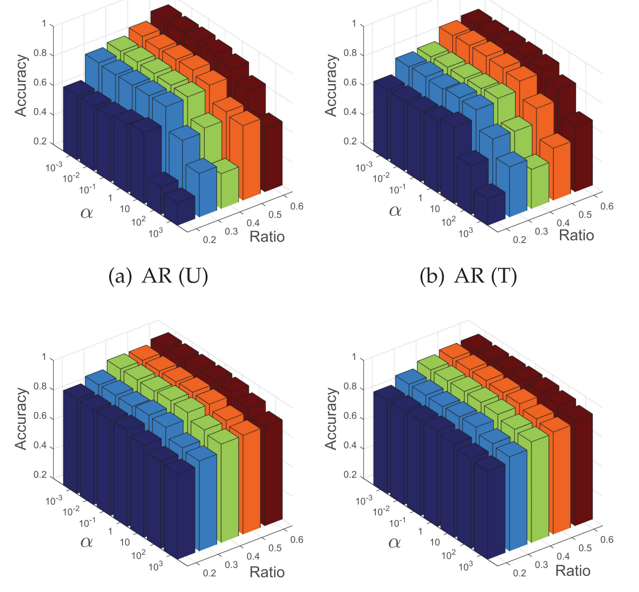
Last, as proposed Algorithm 1 can only obtain a local optimum, we test the sensitivity of our SALOE model in terms of two initializations, P and S. We first repeatedly run our SALOE algorithm 1,000 times with random matrix, and plot the errorbar curve in Fig. 7. Then, we also test the performance of zero matrix and adjacency matrix initialization which calculated as



(a) Coil20 (U) (b) AR (U) (c) MSRA25 (U) (d) YaleB(U) (e) PIE(U)

(f) Coil20 (T) (g) AR (T) (h) MSRA25 (T) (i) YaleB(T) (j) PIE(T)

Fig. 5. Sensitivity analysis for unlabel data (first row) and test data (second row) on regularization parameter α and the number of neighbors k_1 with 20 percent labeled data in training sets.



(a) AR (U) (b) AR (T)

(c) YaleB (U) (d) YaleB (T)

Fig. 6. Sensitivity analysis of regularization parameter α versus ratio of labeled and unlabeled data in training sets.

Eqs. (11) and (12). From the results shown in Fig. 7, we can infer that our algorithm indeed is prone to obtain local optimum, especially when the initialization is random matrix or zero matrix. However, it is still possible to achieve the global optimum, as long as we can find an ideal initialization.

5 Conclusion

In this paper, we first propose a novel semi-supervised adaptive local embedding framework. Subsequently, two locality preserved semi-supervised dimensionality reduction algorithms with whitening and orthogonal constraints, namely SALWE and SALOE respectively are developed. In proposed methods, a local discriminative embedding is learned by using labeled data. Then, projecting all data points into learned embedding space to explore the sub-manifold structure of all samples. Last, the label information can be propagated by many extracted sub-manifold structure. Extensive experiments conducted on two synthetic data sets and several real-world image data sets for image classification, which demonstrates the effectiveness of proposed method.

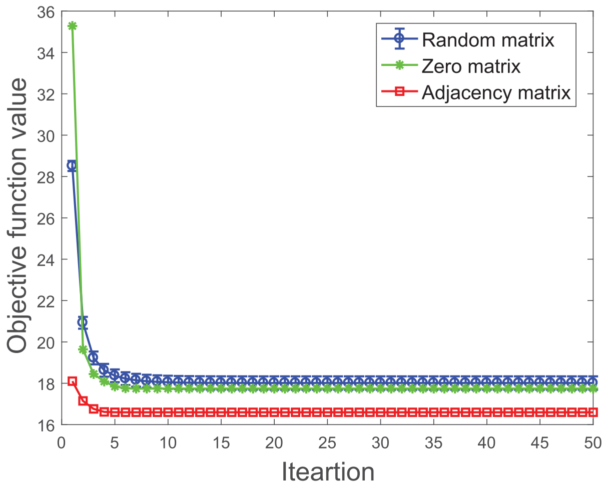


Fig. 7. The curve of objective function value in Eq. (9) with different initializations of P and S in the Algorithm 1.

However, as we described in previous sections, the performance of proposed methods are undesirable when the number of labeled data is small in training. As a result, we will consider exploiting another effective algorithm that can extract much more discriminant information from the small size samples in future works.

Acknowledgment

This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0101902, in part by the National Natural Science Foundation of China under Grant U1803263, Grant 61936014, Grant 61772427 and Grant 61751202, and in part by the Fundamental Research Funds for the Central Universities under Grant G2019KY0501 and Grant 3102019PJ006.

REFERENCES

[1] D. Hu, Z. Wang, H. Xiong, D. Wang, F. Nie, and D. Dou, "Curriculum audiovisual learning," 2020, *arXiv:2001.09414*.

[2] J. Gao, Y. Yuan, and Q. Wang, "Feature-aware adaptation and density alignment for crowd counting in video surveillance," *IEEE Trans. Cybern.*, early access, Dec. 01, 2020, doi: 10.1109/TCYB.2020.3034316.

[3] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning," *IEEE Trans. Neural Netw.*, vol. 20, no. 3, p. 542, 2009.

[4] T. Han, J. Gao, Y. Yuan, and Q. Wang, "Unsupervised semantic aggregation and deformable template matching for semi-supervised learning," 2020, *arXiv:2010.05517*.

[5] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 11th Annu. Conf. Comput. Learn. Theory*, 1998, pp. 92–100.

[6] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proc. Int. Conf. Mach. Learn.*, 1999, pp. 200–209.

[7] X. Zhu, J. Lafferty, and R. Rosenfeld, "Semi-supervised learning with graphs," PhD dissertation, Lang. Technol. Inst., School Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, 2005.

[8] X. J. Zhu and Z. B. Ghahramani, "Learning from labeled and unlabeled data with label propagation," Tech. Rep., CMU-CALD02-107, Carnegie Mellon University, 2002.

[9] X. Zhu, Z. Ghahramani, and J. D. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *Proc. 20th Int. Conf. Mach. Learn.*, 2003, pp. 912–919.

[10] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2004, pp. 321–328.

[11] D. Zhou and B. Schölkopf, "Learning from labeled and unlabeled data using random walks," in *Proc. Joint Pattern Recognit. Symp.*, 2004, pp. 237–244.

[12] M. Tang, F. Nie, and R. Jain, "A graph regularized dimension reduction method for out-of-sample data," *Neurocomputing*, vol. 225, pp. 58–63, 2017.

[13] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, no. Nov., pp. 2399–2434, 2006.

[14] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1548–1560, Aug. 2011.

[15] M. Zheng *et al.*, "Graph regularized sparse coding for image representation," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1327–1336, May 2011.

[16] M. Yin, J. Gao, and Z. Lin, "Laplacian regularized low-rank representation and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 504–517, Mar. 2016.

[17] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[18] M. Yin, S. Xie, Z. Wu, Y. Zhang, and J. Gao, "Subspace clustering via learning an adaptive low-rank graph," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3716–3728, Aug. 2018.

[19] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[20] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2002, pp. 585–591.

[21] M. Sugiyama, "Local fisher discriminant analysis for supervised dimensionality reduction," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 905–912.

[22] X. Li, M. Chen, F. Nie, and Q. Wang, "Locality adaptive discriminant analysis," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 2201–2207.

[23] F. Nie, S. Xiang, and C. Zhang, "Neighborhood minmax projections," in *Proc. Int. Joint Conf. Artif. Intell.*, 2007, pp. 993–998.

[24] D. Cai, X. He, K. Zhou, J. Han, and H. Bao, "Locality sensitive discriminant analysis," in *Proc. Int. Joint Conf. Artif. Intell.*, 2008, pp. 708–713.

[25] T. Zhang, A. Popescul, and B. Dom, "Linear prediction models with graph regularization for web-page categorization," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2006, pp. 821–826.

[26] F. Nie, D. Xu, I. W.-H. Tsang, and C. Zhang, "Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction," *IEEE Trans. Image Process.*, vol. 19, no. 7, pp. 1921–1932, Jul. 2010.

[27] F. Nie, H. Wang, H. Huang, and C. H. Ding, "Adaptive loss minimization for semi-supervised elastic embedding," in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 1565–1571.

[28] D. Wang, F. Nie, and H. Huang, "Large-scale adaptive semi-supervised learning via unified inductive and transductive model," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2014, pp. 482–491.

[29] H. Liu, J. Han, and F. Nie, "Semi-supervised orthogonal graph embedding with recursive projections," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 2308–2314.

[30] K. Xiong, F. Nie, and J. Han, "Linear manifold regularization with adaptive graph for semi-supervised dimensionality reduction," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 3147–3153.

[31] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–7.

[32] Y. Huang, D. Xu, and F. Nie, "Semi-supervised dimension reduction using trace ratio criterion," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 3, pp. 519–526, Mar. 2012.

[33] F. Nie, S. Xiang, Y. Jia, and C. Zhang, "Semi-supervised orthogonal discriminant analysis via label propagation," *Pattern Recognit.*, vol. 42, no. 11, pp. 2615–2627, 2009.

[34] M. Sugiyama, T. Idé, S. Nakajima, and J. Sese, "Semi-supervised local fisher discriminant analysis for dimensionality reduction," *Mach. Learn.*, vol. 78, no. 1, pp. 35–61, 2010.

[35] I. T. Jolliffe, "Principal component analysis and factor analysis," in *Principal Component Analysis*. Berlin, Germany: Springer, 1986, pp. 115–128.

[36] Y. Zhang and D.-Y. Yeung, "Semi-supervised discriminant analysis using robust path-based similarity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[37] Y. Zhou and S. Sun, "Semisupervised tangent space discriminant analysis," *Math. Problems Eng.*, vol. 2015, 2015, Art. no. 706180.

[38] H. Zhao, Z. Wang, and F. Nie, "A new formulation of linear discriminant analysis for robust dimensionality reduction," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 4, pp. 629–640, Apr. 2019.

[39] Z. Wang, F. Nie, C. Zhang, R. Wang, and X. Li, "Capped ℓ_p-norm LDA for outliers robust dimension reduction," *IEEE Signal Process. Lett.*, vol. 27, pp. 1315–1319, Jul. 2020.

[40] Z. Wang, F. Nie, L. Tian, R. Wang, and X. Li, "Discriminative feature selection via a structured sparse subspace learning module," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, 2020, pp. 3009–3015.

[41] P. Indyk and R. Motwani, "Approximate nearest neighbors: Towards removing the curse of dimensionality," in *Proc. 13th Annu. ACM Symp. Theory Comput.*, 1998, pp. 604–613.

[42] F. Nie, Z. Wang, R. Wang, Z. Wang, and X. Li, "Towards robust discriminative projections learning via non-greedy $l_{2,1}$-norm MinMax," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Dec. 24, 2019, doi: 10.1109/TPAMI.2019.2961877.

[43] F. Nie, Z. Wang, L. Tian, R. Wang, and X. Li, "Subspace sparse discriminative feature selection," *IEEE Trans. Cybern.*, early access, Oct. 15, 2020, doi: 10.1109/TCYB.2020.3025205.

[44] F. Nie, Z. Wang, R. Wang, Z. Wang, and X. Li, "Adaptive local linear discriminant analysis," *ACM Trans. Knowl. Discov. Data*, vol. 14, no. 1, pp. 1–19, 2020.

[45] F. Nie, Z. Wang, R. Wang, and X. Li, "Submanifold-preserving discriminant analysis with an auto-optimized graph," *IEEE Trans. Cybern.*, vol. 50, no. 8, pp. 3682–3695, Aug. 2020.

[46] Z. Wang, F. Nie, R. Wang, H. Yang, and X. Li, "Local structured feature learning with dynamic maximum entropy graph," *Pattern Recognit.*, vol. 111, 2021, Art. no. 107673. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320320304763

[47] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[48] E. Kokiopoulou and Y. Saad, "Orthogonal neighborhood preserving projections: A projection-based dimensionality reduction technique," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2143–2156, Dec. 2007.

[49] D. Cai, X. He, J. Han, and H.-J. Zhang, "Orthogonal laplacianfaces for face recognition," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3608–3614, Nov. 2006.

[50] H. Zhao, Z. Wang, and F. Nie, "Orthogonal least squares regression for feature extraction," *Neurocomputing*, vol. 216, pp. 200–207, 2016.

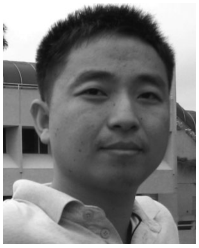
[51] S. A. Nene, K. N. Shree, and M. Hiroshi, "Columbia object image library (COIL-20)," Tech. Rep. CUCS-005-96, 1996.

[52] X. He, S. Yan, Y. Hu, and H.-J. Zhang, "Learning a locality preserving subspace for visual recognition," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, 2003, pp. 385–392.

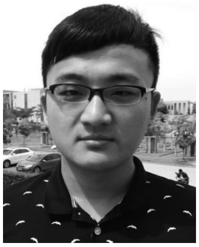
[53] A. Martinez and R. Benavente, "The AR face database," CVC Tech. Rep. no. 24, 1998.

[54] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.

[55] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression (PIE) database," in *Proc. 5th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2002, pp. 53–58.



Feiping Nie received the PhD degree in computer science from Tsinghua University, Beijing, China, in 2009, and is currently a full professor with Northwestern Polytechnical University, China. His research interests include machine learning and its applications, such as pattern recognition, data mining, computer vision, image processing, and information retrieval. He has published more than 100 papers in the following journals and conferences: the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *International Journal of Computer Vision*, *IEEE Transactions on Image Processing*, *IEEE Transactions on Neural Networks and Learning Systems*, *IEEE Transactions on Knowledge and Data Engineering*, ICML, NIPS, KDD, IJCAI, AAAI, ICCV, CVPR, and ACM MM. His papers have been cited more than 10000 times and the H-index is 57. He is now serving as an associate editor or PC member for several prestigious journals and conferences in the related fields.



Zheng Wang received the MS degree from the Anhui University, Hefei, China, and is currently working toward the PhD degree in the School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. His research interests mainly focus on dimensionality reduction, multimodal machine learning and its applications. He has published several papers in journals such as the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Cybernetics*, *IEEE Transactions on Knowledge and Data Engineering*, *ACM Transactions on Knowledge Discovery from Data*, and *Pattern Recognition*.

Rong Wang (Member, IEEE) received the BE degree in information engineering, the ME degree in signal and information processing, and the PhD degree in computer science from the Xi'an Research Institute of Hi-Tech, Xi'an, China, in 2004, 2007, and 2013, respectively. From 2007 to 2013, he was also with the Department of Automation, Tsinghua University, Beijing, China, for the PhD degree. His current research interests include machine learning and signal processing, together with their applications including pattern recognition, image processing, and computer vision.

Xuelong Li (Fellow, IEEE) is a full professor with the School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, P.R. China.