Semi-supervised Sparse Subspace Clustering Based on Re-weighting

Qiaoyan Li, Xue Zhao and Hengdong Zhu

***Abstract*—The traditional sparse subspace clustering algorithms are easily affected by the similarity matrix, which may lead to different clustering results by different similarity matrix. That is to say, constructing a reasonable similarity matrix is the key to sparse subspace clustering. Based on re-weighting subspace clustering, a semi-supervised sparse subspace clustering algorithm based on re-weighting is proposed in the paper. Firstly, the global similarity structure of the data can be better captured by constraining the coefficient between the cannot-link labels to be 0. Secondly, with the help of re-weighted l_1-norm minimization sparse optimization framework, the adaptive similarity matrix can be obtained. Furthermore, the above similarity matrix can be adjusted by using prior information. Experimental results indicate that the proposed clustering algorithm is more efficient than other clustering algorithms on benchmark data sets.**

***Index Terms*—pairwise constraint, re-weighting, semi-supervised, sparse subspace clustering.**

I. Introduction

IN today's society, all kinds of data are flooded in people's lives. High-dimensional data widely is available in machine learning, signal and image processing, computer vision, pattern recognition and other fields. Clustering analysis of high-dimensional data will enhance the calculation time and storage requirements of the algorithm. Moreover, when high-dimensional data contains noise, traditional clustering algorithms, such as K-means clustering, spectral clustering , fuzzy clustering and so on, can no longer handle this type of data well. However, considering that high-dimensional data are often distributed on the union of multiple low-dimensional subspaces, obtaining the low-dimensional structure of high-dimensional data can not only reduce the computational cost and storage requirements of the algorithm, but also reduce the noise influence in the data and improve the performance of clustering analysis [1]. Therefore, the subspace clustering problem is proposed. Given a set of points drawn from a union of subspaces, the task is to find the number of subspaces, their dimensions, the basis for each subspace, and the segmentation of the data[2]. After recent years of development and research, many subspace clustering methods have been proposed, which can be roughly divided into five categories: based on matrix factorization[3], based on algebra [4], [5], based on iteration [6], [7], based on statistics [8], [9] and based on spectral clustering [10], [11].

Sparse subspace clustering (SSC) is a subspace clustering method based on spectral clustering. The algorithm has attracted widespread attention, and its main idea is based on the fact that each point in a union of subspaces has a sparse representation with respect to a dictionary formed by all other data points [12]. Therefore, the core problem is to acquire the appropriate coefficient matrix, and then use the similarity matrix by spectral clustering to obtain the final clustering results[13]. In order to get a more "appropriate" coefficient matrix, various restrictions and constraints are usually added to the coefficient matrix in the algorithm model to force it to have an ideal block diagonal structure. This ideal structure can help discover and mine the potential manifold structure information behind the data for clustering.

In [14], a weighted sparse optimization framework is proposed by using the spatial geometry of data points to weight the representation coefficients in the sparse optimization framework. The re-weighted l_1 norm was used to replace the traditional l_1 norm, and a re-weighted sparse subspace clustering (RSSC) algorithm was proposed[15]. A structured sparse subspace clustering (SSSC) algorithm is proposed, which uses a unified optimization framework to automatically combine the coefficient matrix and spectral clustering[16]. The algorithm is built on expressing each data point as a structured sparse linear combination of all other data points, where the structure is induced by a norm that depends on the unknown segmentation. A scalable sparse subspace clustering by orthogonal matching pursuit (SSSC-OMP) algorithm based on orthogonal matching pursuit was proposed to solve the sparse solution of sparse subspace model[17]. In order to reduce the computational complexity of using the OMP method for large-scale data, a learning OMP (learning orthogonal matching pursuit, LOMP) algorithm is proposed in [18]. An improved SSC algorithm is used to select the appropriate band subset to solve the problem of hyperspectral image classification in [19].

The above are all unsupervised clustering algorithms. In practical applications, a priori information of a small amount of data can often be obtained. If we just use the unsupervised learning, these priori information will be wasted. In order to make full use of the prior information to improve the clustering accuracy, semi-supervised learning is proposed. According to the different forms of supervision information, semi-supervised clustering can be divided into: semi-supervised clustering based on label information and semi-supervised clustering based on pairwise constraints[20]. In [21], a unified manifold learning framework (FME) for semi-supervised was proposed and further studies can be found in [22]. By using a novel adaptive loss minimization method, a semi-supervised elastic embedding (SEE) algorithm were

Manuscript received July 29, 2022; revised December 6, 2022. This work is supported by the Natural Science Foundation of China (61976130), Natural Science Foundation of Shaanxi Province (2020JQ-923), key research and development projects of Shaanxi Province (2018KW-021).

Qiaoyan Li is an associate professor of School of Science, Xi'an Polytechnic University, Xi'an, Shaanxi Province, 710048, China (email: liqiaoyan@xpu.edu.cn).

Xue Zhao is a postgraduate student of Xi'an Polytechnic University, Xi'an City, Shaanxi Province, 710048, China (email: 872421556@qq.com).

Hengdong Zhu is a postgraduate student of Xi'an Polytechnic University, Xi'an City, Shaanxi Province, 710048, China (email: ZhuHengDong1997@163.com).

proposed in [23]. In order to use the neighboring information, a semi-supervised sparse subspace clustering algorithm LR-LDS^{4}C was given in [24]. For making better use of the pairwise constraints of the given labels, a label-guided weighted semi-supervised neutrosophic clustering algorithm LG-WSSNCM was introduced in [25].

Semi-supervised sparse subspace clustering is an important semi-supervised learning method [26]. How to make full use of the supervised information to improve the results of sparse subspace clustering is an important problem. In the paper, the label information of some samples is selected as the supervision information, and a re-weighted semi-supervised sparse subspace clustering algorithm (SSRSSC) is proposed based on the re-weighted l_1 norm minimization framework. The contributions of the given algorithm are not only to consider combining the label information with the construction of the similarity matrix, but also make full use of the supervision information to cluster. Specifically, firstly, with the help of pairwise constraint information and re-weighted l_1 norm minimization sparse optimization framework, the optimal coefficient matrix Z is obtained by updating iteration; secondly, similarity matrix is constructed by using Z, and the clustering model is adjusted by Laplacian regularization term of the constructed similarity matrix to obtain the projected data F. Finally, the final clustering result is obtained by taking the largest of each row of F.

The remainder of this paper is organized as following. In Section II, we introduce related work. In Section III, we introduced the establishment and solution of the SSRSSC model. Experimental results and analysis are presented in Section IV. Finally, Section V concludes our paper.

II. Related work

A. Sparse subspace clustering algorithm

The analysis of high-dimensional data can be transformed into the study of coefficient matrix of high-dimensional data in low-dimensional space. Then, for the input high-dimensional data matrix $X = [x_1, \cdots, x_n] \in \mathbb{R}^{d\times n}$, it can be expressed $X = XZ$, where $X \in \mathbb{R}^{d\times n}$ as a dictionary and $Z \in \mathbb{R}^{n\times n}$ as a coefficient matrix. $Z_{ij} = 0$ indicates that the sample points x_i and x_j belong to different subspaces. Arrange the input data column by column according to its subspace category. Ideally, Z should have a block diagonal structure[13].

$$Z = \begin{pmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_k \end{pmatrix},$$

where Z_i is the representation coefficient matrix of the data in the i-th subspace. When Z with the block diagonal structure is obtained, the subspace structure of input data can be obtained. The sparse subspace clustering algorithm is to use different norm constraints Z to make it have the ideal structure of block diagonal as far as possible, so as to explore the subspace structure of data. Therefore, the basic model of sparse subspace can be expressed as follows:

$$\begin{aligned} &\min_{Z} f(Z) \\ &s.t. \;\; X = XZ, Z_{ii} = 0. \end{aligned} \tag{1}$$

where $f(Z)$ is a sparse function of the self-representation matrix Z , such as $\|Z\|_1$ and $\|Z\|_{2,1}$. The constraint $Z_{ii} = 0$ is used to avoid the special case where each data is represented only by itself. When noise is considered in the model, the constraint condition $X = XZ$ can be replaced by $X = XZ + E$, and the appropriate regular term constraint can be applied to noise E. Elhamifar and Vidal proposed the most classical sparse subspace clustering model as follows:

$$\begin{aligned} &\min_{Z,E} \|Z\|_1 + \lambda\|E\|_{2,1} \\ &s.t. \;\; X = XZ + E, Z_{ii} = 0. \end{aligned} \tag{2}$$

where $\|Z\|_1$ represents the l_1 norm of matrix Z, which has strong sparsity[13]; $E \in \mathbb{R}^{d\times n}$ is a noise matrix, $\|E\|_{2,1}$ represents the $l_{2,1}$ norm of matrix E, which has strong robustness [27]; by introducing a robust error term $\|E\|_{2,1}$, the influence of outliers and noise points on the robust segmentation results of SSC can be reduced; the parameter $\lambda > 0$ is used to balance the two terms in the objective function.

B. Re-weighted sparse optimization framework

The SSC algorithm use the sparse representation of vectors lying on a union of subspaces to cluster the data into separate subspaces. In order to obtain the sparse representation of each data point, the re-weighted l_1 norm minimization is used to perform convex relaxation. At the same time, in practical problems, data points are often mixed with sparse singular values and noise. In addition, the data are often distributed on the union of affine subspaces rather than linear subspaces. So a re-weighted sparse optimization framework is established as follows:

$$\begin{aligned} &\min_{Z,E} \|W \odot Z\|_1 + \lambda\|E\|_{2,1} \\ &s.t. \;\; X = XZ + E, Z_{ii} = 0. \end{aligned} \tag{3}$$

where $W \in \mathbb{R}^{n\times n}$ is a re-weighted diagonal matrix. In reference[16], $f_{log}(x)$ function is introduced, and its re-weighted matrix updating formula is as follows $W_i^k = \frac{1}{|x_i^k|+\epsilon}$.

However, in practical applications, subspaces are interdependent and data may be in nonlinear manifolds. In these cases, the block diagonal structure of Z is often greatly affected. In order to overcome this difficulty, in the second section, a new method is proposed to extend the re-weighted sparse optimization model (SSRSSC) in RSSC framework by considering the label information of samples.

III. Establishment and solution of objective function

A. The establishment of objective function

In order to make Z have a better block diagonal structure, the label information of known samples is integrated into the original RSSC framework, and a new re-weighted semi-supervised sparse representation (SSRSSC) model is

proposed. The core idea of SSRSSC is to know the prior information of some samples when solving the RSSC problem so that the samples on the same subspace should be divided into a cluster. Therefore, Z should be kept sparse and have block diagonal structure. Since the labels of some samples are known, the coefficient $Z_{i,j}$ between two data points from different clusters is directly forced to zero. Therefore, the SSRSSC model solves the following problems:

$$\begin{aligned} &\min_{Z,E} \|W \odot Z\|_1 + \lambda \|E\|_{2,1} \\ &s.t. \;\; X = XZ + E, Z_{i,j} = 0, \forall (i,j) \subset \Omega, Z^T \mathbf{1} = \mathbf{1}. \end{aligned} \tag{4}$$

where $\mathbf{1}$ is an all-one vector, Ω the set of edges between two labeled samples from different classes and sample itself, whereby $(i,j) \subset \Omega$ indicates that sample x_i and sample x_j are not in the same class. As we can see from (1), similar to SSC, SSRSSC also seeks the sparse representation Z among all the data points. Meanwhile, by enforcing $Z_{i,j} = 0$, $(i,j) \subset \Omega$, it makes use of the label information to help prevent the block-diagonal structure of Z from being destroyed in real world scenarios. By enforcing the sum-to-one constraint on the rows of the coefficient matrix, we hope to obtain the invariance to translations. The same trick is also used by existing methods, such as the popular Locally Linear Embedding (LLE)[28]. Since the SSRSSC problem (4) is convex, it can be efficiently solved by fast first-order optimization methods, as we describe next.

B. The solution of objective function

Then, applying LADMAP[29] to the standard form (4) yields the following updating rules. The problem (4) is solved by using the augmented Lagrange multiplier method as follows:

$$\begin{aligned} \text{Ł}(Z,E,Y_1,y_2,y_3,\beta) &= \|W \odot Z\|_1 + \lambda \|E\|_{2,1} \\ &+ < Y_1, XZ + E - X > + < y_2, \rho_\Omega(Z) > \\ &+ < y_3, Z^T \mathbf{1} - \mathbf{1} > + \frac{\beta}{2}(\|XZ + E - X\|_F^2 \\ &+ \|\rho_\Omega(Z)\|_2^2 + \|Z^T \mathbf{1} - \mathbf{1}\|_2^2) \end{aligned} \tag{5}$$

a) Updating Z, fix E, Y_1, y_2, y_3, β: Here, we first note that

$$\begin{aligned} F_Z = &< Y_1, XZ + E - X > + < y_2, \rho_\Omega(Z) > \\ &+ < y_3, Z^T \mathbf{1} - \mathbf{1} > + \frac{\beta}{2}(\|XZ + E - X\|_F^2 \\ &+ \|\rho_\Omega(Z)\|_2^2 + \|Z^T \mathbf{1} - \mathbf{1}\|_2^2) \end{aligned} \tag{6}$$

The derivation of formula (6) is as follows:

$$\begin{aligned} \nabla F_Z = &X^T Y_1 + \beta(X^T(XZ + E - X)) \\ &+ \rho_\Omega^*(y_2) + \mathbf{1}(y_3)^T + \rho_\Omega^*(\rho_\Omega(Z)) + \mathbf{1}\mathbf{1}^T Z - \mathbf{1}\mathbf{1}^T \end{aligned} \tag{7}$$

For any Z_1, Z_2 the following formula (7) is as follows:

$$\begin{aligned} &\beta(\|X^T X(Z_1 - Z_2) + \rho_\Omega^*(\rho_\Omega(Z_1 - Z_2)) \\ &\quad + \mathbf{1}\mathbf{1}^T(Z_1 - Z_2)\|_F) \\ &\le \beta(\|X^T X(Z_1 - Z_2)\|_F + \|\rho_\Omega^*(\rho_\Omega(Z_1 - Z_2))\|_F \\ &\quad + \|\mathbf{1}\mathbf{1}^T(Z_1 - Z_2)\|_F) \\ &\le \beta(\|X\|_2^2 + n + 1)\|(Z_1 - Z_2)\|_F \end{aligned} \tag{8}$$

Here,$\|X\|_2^2 + n + 1 = \eta_A$ according to the above formula, the Lipschitz constant L_F^Z is $\beta(\|X\|_2^2 + n + 1) = \beta\eta_A$.

$$\|\nabla F(Z_1) - \nabla F(Z_2)\|_F \le L_F^Z \|Z_1 - Z_2\|_F \tag{9}$$

According to Taylor's formula, formula (5) can be solved by the following formula (10):

$$\begin{aligned} L(Z) = &\|W \odot Z\|_1 + < Z - Z_K, \nabla F(Z_K) > \\ &+ \frac{L_F^Z}{2}\|Z - Z_K\|_F^2 \end{aligned} \tag{10}$$

Formulate the above formula, it is as follows:

$$\|W \odot Z\|_1 + \frac{L_F^Z}{2}\|Z - (Z_k - \frac{1}{L_F^Z}\nabla F(Z_K))\|_F^2 \tag{11}$$

According to the proximity point operator, it is as follows:

$$\begin{aligned} Z_{k+1} &= S_{\frac{W}{L_F^Z}} Z_k - \frac{1}{L_F^Z}\nabla F(Z_k) \\ &= S_{\frac{W}{\beta\eta_A}} Z_k - \frac{1}{L_F^Z}\nabla F(Z_k) \end{aligned} \tag{12}$$

Here, we note that $Z_K - \frac{1}{L_F^Z}\nabla F(Z_K) = Z_K^*$.

$$Z_{k+1} = sgn(Z_k^*)\max(|Z_k^*| - \frac{W}{\beta\eta_A}, 0) \tag{13}$$

b) Updating E, fix Z, Y_1, y_2, y_3, β: Similarly, we note that

$$F_E = < Y_1, XZ + E - X > + \frac{\beta}{2}(\|XZ + E - X\|_F^2) \tag{14}$$

According to references[21], it is as follows:

$$\begin{aligned} E_{k+1} &= \max(1 - \frac{\lambda}{L_F^E(\|E^*\|_2)}, 0)E^* \\ &= \max(1 - \frac{\lambda}{\beta(\|E^*\|_2)}, 0)E^* \end{aligned} \tag{15}$$

Here, $E_k^* = E_k - \frac{1}{L_F^E}\nabla F(E_k)$.

c) Updating Y_1,y_2,y_3, fix Z, E, β:: Third, the Lagrange multiplier Y_1,y_2,y_3 is updated as:

$$Y_{1,k+1} = Y_{1,k} + \beta(XZ + E - X) \tag{16}$$

$$y_{2,k+1} = y_{2,k} + \beta(\rho_\Omega(Z)) \tag{17}$$

$$y_{3,k+1} = y_{3,k} + \beta(Z^T \mathbf{1} - \mathbf{1}) \tag{18}$$

d) Updating β , fix Z, E, Y_1, y_2, y_3: the penalty β is updated adaptively as follows:

$$\beta_{k+1} = \min(\beta_{max}, \gamma\beta_k) \tag{19}$$

Where

$$\gamma = \begin{cases} \gamma_0, & ZE \le \varepsilon \\ 1, & \text{otherwise.} \end{cases} \tag{20}$$

where $\gamma_0 \geq 1$ is a constant, $0 < \varepsilon < 1$ is a threshold and $ZE = \beta_k \max(\sqrt{\eta_A}\|Z_{k+1} - Z_k\|_F, \sqrt{\eta_B}\|E_{k+1} - E_k\|_F)$.

Based on the above analysis, the detailed procedure for the given algorithm is summarized in Algorithm 1. We adopt the popular Local and Global Consistency (LGC)[30] as the classification framework. Specifically, LGC builds upon an undirected graph, and utilizes the graph and known labels to recovery a continuous classification function $F \in \mathbb{R}^{n\times c}$ by optimizing the following energy function:

$$\min_{F\in\mathbb{R}^{n\times c}} tr(F^T L_Z F + \mu(F-Y)^T(F-Y)) \tag{21}$$

where c is the number of classes, $Y \in \mathbb{R}^{n\times c}$ is the label matrix, in which $Y_{i,j} = 1$ if sample x_i is associated with label j for $j \subset 1, 2, \cdots, c$, and $Y_{i,j} = 0$ otherwise. L_Z is the normalized graph Laplaction $L_Z = D^{-\frac{1}{2}}(D - Z)D^{-\frac{1}{2}}$, in which D is a diagonal matrix with $D_{i,j} = \sum_j Z_{i,j}$. The weight $\mu \subset [0, \infty)$ balance the local fitting and global smoothness of the function F.

TABLE I: SSRSSC algorithm

Algorithm 1 LADMAP for Solving the SSRSSC Problem

Input: Data matrux $X = [x_1, \cdots, x_n] \in \mathbb{R}^{d\times n}$,
balance parameter $\lambda, \eta_A, \eta_B, \mu$, and indices set Ω.
Initialize $Z = 0$, $E = 0$, $Y_1 = 0$, $y_2 = 0$, $y_3 = 0$,
$\beta_{max} = 10^{10}$, $\gamma = 1.1$, $\varepsilon = 10^{-8}$
Do
Update Z by(13).
Update E by(15).
Update Y_1 by(16).
Update y_2 by(17).
Update y_3 by(18).
Update β by(19).
While
$\|XZ + E - X\|_\infty < \varepsilon$ and
$\beta_k \max(\sqrt{\eta_A}\|Z_{k+1} - Z_k\|_F, \sqrt{\eta_B}\|E_{k+1} - E_k\|_F) < \varepsilon$.
Output: The optimal solution Z.

According to the optimization process of Algorithm 1, a theoretical complexity analysis of the SSRSSC algorithm will be performed. Denote n, d, c and t are the number of samples, the data's dimension, the number of clusters and the maximum number of iterations respectively. The complexity of the SSRSSC algorithm has three main components. (1) Construction of constraint set Ω: Considering that Ω is related to the number of manually labeled samples, the complexity of this part does not exceed $O(n^2)$. (2) The alternating iterative process of variables Z, E, Y_1, y_2, y_3 and β in the SSRSSC algorithm: where the complexity corresponding to variables Z, E, Y_1, y_3 and β is $O[t(n^2d+n^2+dn)]$. Assuming that the number of manually labeled samples is m, the complexity corresponding to variable y_2 is $O(tm^2)$. Thus, the total complexity of this part is $O[t(n^2d+n^2+dn+m^2)]$. (3) The clustering process: where the complexity corresponding to the variable F is $O(n^3)$. It is worth noting that we did not choose a discretization procedure such as K-means to process F and obtain the clustering result. The most valuable F_{ij} (maximum value) corresponds to the j-th cluster. Therefore, the total complexity of this part is $O(n^3)$. In summary, the total complexity of the given method is $O[t(n^2d + n^2 + dn + m^2) + n^3 + n^2]$.

IV. Experiments

A. Datasets

To illustrate the effectiveness of the given algorithm, we will conduct a series of experiments, and we performed experiments on four benchmark datasets: 1) Isolet dataset; 2) Yale dataset; 3) Coil 20 dataset; 4) USPS dataset.

1) Isolet dataset

The number of categories of the Isolet data set is 26 , the number of sample points is 1560 , and the dimension is 617.

2) Yale dataset

The Yale Faces dataset contains 165 grayscale images in GIF format of 15 individuals. There are 11 images per subject, one per different facial expression or configuration. In the experiments, the original images were normalized and cropped into 32×32 pixels for clustering.

3) Coil 20 dataset

Evaluation on Visual Object Recognition: We verify the importance of label information for graph learning for non-linear manifolds by conducting visual object recognition experiments on the COIL 20 dataset. The dataset contains 20 objects. The images of each objects were taken 5 degrees apart as the object is rotated on a turntable, resulting in 72 images for each object. The size of the grayscale image is 3232 pixels.

TABLE II: Clustering performance on the Isolet dataset.

	LRR	SR	SSLRR	SSRSSC
0.1	23.23±0.03	27.36±0.07	21.84±0	16.91±0.004
0.2	16.23±0.01	17.36±0.02	15.11±0.01	12.23±0.02
0.3	13.54±0.01	11.63±0.01	11.26±0.1	9.77±0.01
0.4	10.64±0.002	8.77±0.01	8.71±0.001	7.88±0.0005
0.5	8.69±0.001	6.43±0.002	6.55±0.002	5.83±0.001
0.6	6.51±0.03	4.88±0.03	4.98±0.001	4.71±0.001

TABLE III: Clustering performance on the Yale dataset.

	LRR	SR	SSLRR	SSRSSC
0.1	19.66±0.02	17.37±0.02	25.63±0.33	13.37±0.003
0.2	9.21±0.003	13.27±0.004	17.14±0.04	9.15±0.001
0.3	5.35±0.002	8.61±0.004	14.65±0.03	5.76±0.003
0.4	3.37±m0.0003	5.54±0.002	11.64±0.006	3.69±0.002
0.5	2.18±0.001	3.62±0.01	9.88±0.02	2.75±0.001
0.6	2.03±0.001	2.83±0.001	7.96±0.004	1.62±0.0004

TABLE IV: Clustering performance on the Coil 20 dataset.

	LRR	SR	SSLRR	SSRSSC
0.1	17.86±0.03	8.92±0.01	19.38±0.02	6.64±0.003
0.2	9.54±0.004	2.78±0.002	11.38±0.01	4.18±0.009
0.3	5.49±0.008	1.65±0.002	7.28±0.01	3.22±0.01
0.4	4.09±0.002	0.91±0.002	5.40±0.002	2.04±0.003
0.5	2.87±0.002	0.78±0.002	3.92±0.003	1.35±0.001
0.6	2.03±0.001	0.44±0.0003	3.01±0.003	0.78±0.002

TABLE V: Clustering performance on the USPS dataset.

	LRR	SR	SSLRR	SSRSSC
0.1	13.69±0.26	7.69±0.31	7.27±0.29	6.10±0.22
0.2	10.98±0.14	5.66±0.27	6.44±0.27	5.65±0.19
0.3	9.33±0.14	4.90±0.28	5.76±0.18	5.00±0.16
0.4	7.85±0.23	3.90±0.15	4.88±0.20	4.14±0.18
0.5	6.95±0.22	3.12±0.14	4.29±0.16	3.45±0.24
0.6	6.17±0.14	2.63±0.12	3.91±0.16	2.65±0.08

4) USPS dataset

The USPS digit dataset is described in [26]. A popular subset contains $9,298$ images of 16×16 pixels recording handwriting digits from 0 to 9.

B. Preliminary analysis of the given algorithm

In this subsection, we will give preliminary analysis of the given the SSRSSC algorithm. Firstly, the given algorithm will be compared with the other three algorithms to evaluate the clustering performance. Secondly, the performance of block diagonal structure will be shown. The prior information of all algorithms is selected according to the proportion of 0.1,0.2,0.3,0.4,0.5,0.5,0.6, and the label information of the corresponding proportion is randomly selected as the prior information in each class of all data sets. According to reference[26], the value of the parameter μ of the above algorithm is 0.9. In this paper, perform 10 independent repeated experiments for each algorithm in proportion to the label information, and the mean and variance were used to represent the final clustering results of the algorithm.

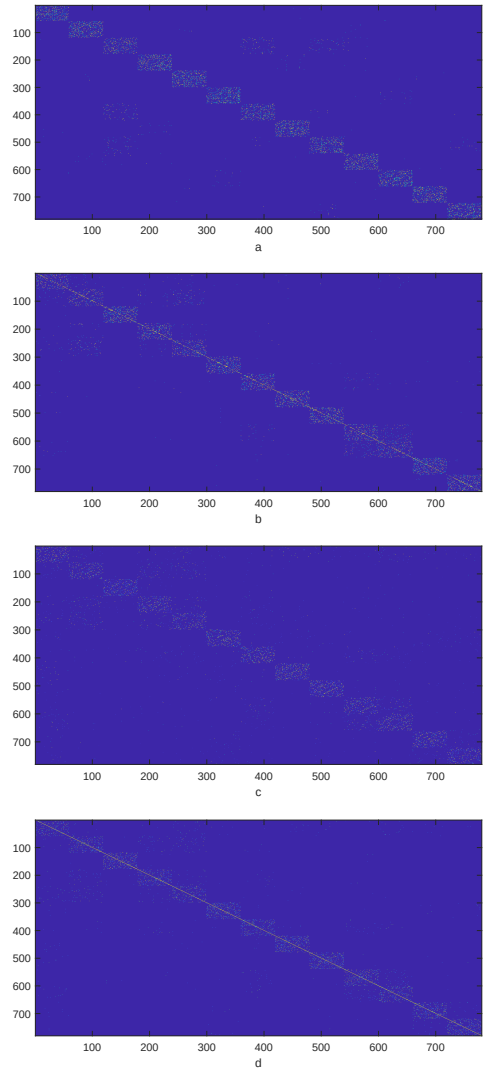

Fig. 1: Intuitive graph of similarity matrix of SSRSSC algorithm(a),SSLRR algorithm(b),SR algorithm(c) and LRR algorithm(d).

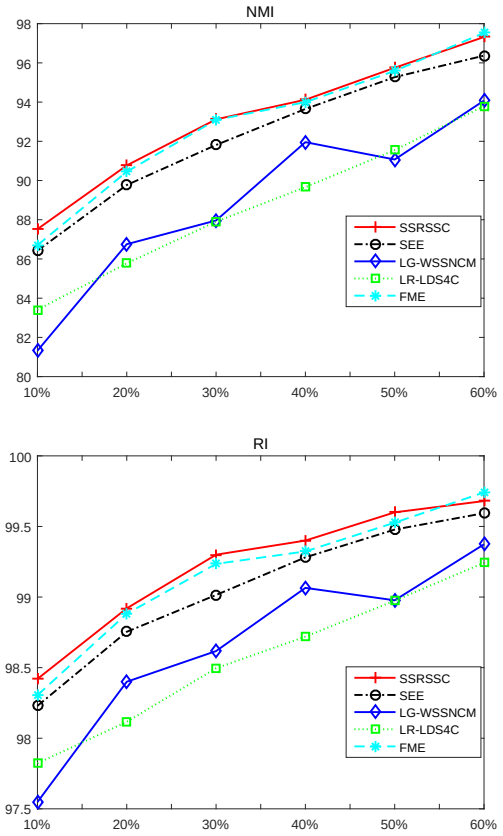

Fig. 2: NMI and RI evaluation results of five different semi-supervised clustering algorithms on Isolet dataset

To demonstrate the performance of the given algorithm, the LRR algorithm[31], SR algorithm[32], and SSLRR algorithm[26] are selected and the clustering error rate will be used to evaluate the clustering performance. Table $2-5$ shows the clustering results of Isolet, Coil, Yale and USPS datasets. From Table 2 and Table 3, we can find that with the increase of label information, the error rate of the SSRSSC algorithm is significantly reduced and is lower than other comparison algorithms. From Table 4 and Table 5, we can find that when the label information is only 0.1, the error rate of the SSRSSC algorithm is the lowest; with the increase of label information, the error rate of the SSRSSC algorithm is only slightly higher than that of the SR algorithm, but still lower than the LRR algorithm and SSLRR algorithm. In summary, the clustering result of the SSRSSC algorithm proposed in this paper is slightly better than other comparison algorithms.

In order to show the block diagonal structure of the given algorithm, we take the first 780 points of the Isolet data set, and when the label information is 30, the intuitive diagram of the similarity matrix of the SSRSSC algorithm and the comparison algorithm is shown in Figure 1. From Figure 1, it is obvious that the SSRSSC algorithm proposed in this paper has a better block diagonal structure than the comparison algorithm.

C. Further performance analysis

In this subsection, We will further analyze the performance of the given SSRSSC algorithm. The given algorithm will

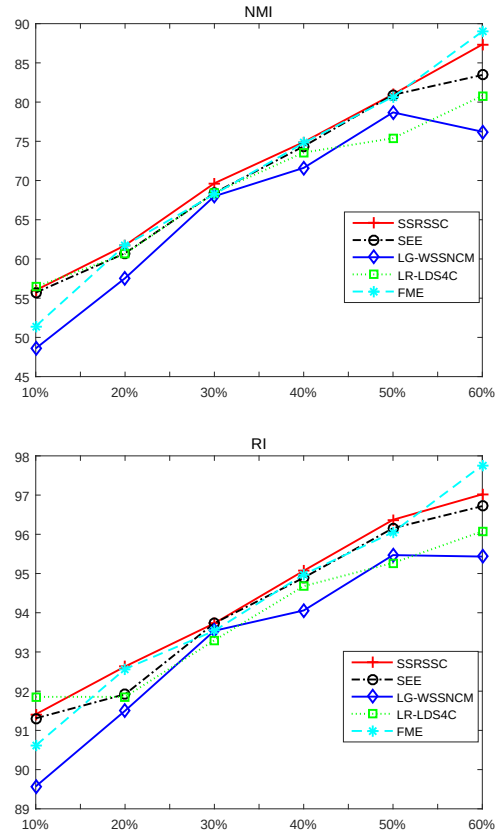

Fig. 3: NMI and RI evaluation results of five different semi-supervised clustering algorithms on Yale dataset

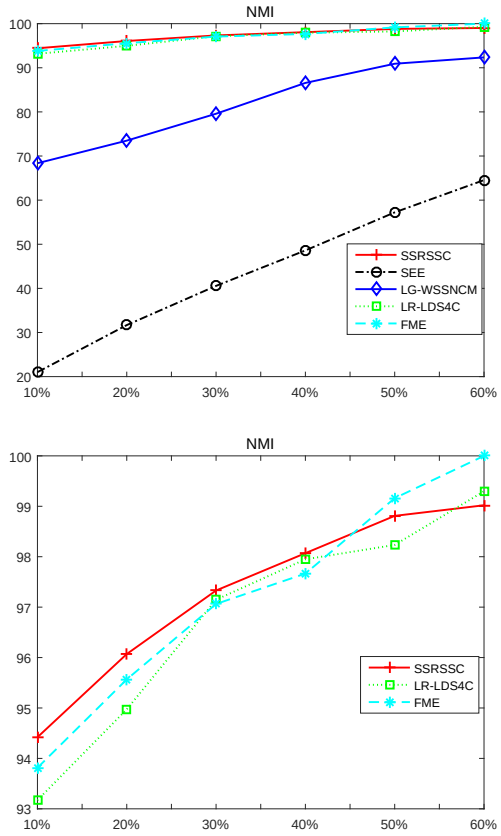

Fig. 4: NMI evaluation results of five different semi-supervised clustering algorithms on Coil dataset, the following figure is a zoom in for the above figure.

be compared with FME [21], SEE [23], LR-LDS^{4}C [24] and LG-WSSNCM [25], and ACC and NMI will be used to evaluate the clustering performance. In the parameter setting, FME: $\mu = [0.001, 0.01, 0.1, 1, 10, 100, 1000], \gamma = [0.001, 0.01, 0.1, 1, 10, 100, 1000]$. SEE: $\gamma = [0.001, 0.01, 0.1, 1, 10, 100, 1000]$, $\sigma = [0.1, 0.5, 1, 1.5, 2, 2.5, 3]$. LR-LDS^{4}C: $k = [5, 10, 20, 30, 40, 45, 50], \lambda = [0.001, 0.01, 0.1, 1, 10, 100, 1000]$ and LG-WSSNCM: $w_1 = 0.3$, $w_2 = 0.4, w_3 = 0.55$. SSRSSC: $\lambda = [0.001, 0.01, 0.1, 1, 10, 100, 1000]$, ϵ select in $[0.1, 3]$. We choose the best experimental results for comparison.

Fig. 2-7 show the evaluation results of the NMI and RI for different semi-supervised clustering algorithms on the given four datasets. Horizontal axis is the prior information and longitudinal axis is the value of NMI or RI. The prior information of all algorithms is selected according to the proportion p is 0.1, 0.2, 0.3, 0.4, 0.5 or 0.6 respectively. It can be obtained that in most cases, the clustering performance of all semi-supervised clustering algorithms will decrease with the decrease of the percentage of labeled samples. When p = 0.6, FME shows better performance compared with the given algorithm SSRSSC. For Coil dataset, the performance of SEE is poor, which possibly due to the given parameters. That is to say, the SEE algorithm is not optimal under the given parameters. The same situation is also reflected in the algorithm LG-WSSNCM for dataset USPS. However, under the same parameter values, the designed experiment can still reflect the overall performance of the algorithm. More importantly, in most cases, the given algorithm SSRSSC is superior to the other four semi-supervised clustering algorithms, which fully show the effectiveness of the given algorithm.

D. Parameters analysis

There are two parameters in the SSRSSC algorithm. Generally, the results of the algorithm will change by the value of parameters. Without losing generality, we select Isolet dataset to reveal the changes by the parameters and the given prior information proportion and the performance of ACC is selected to show the results. Fig.6 shows the ACC values that vary with the parameters λ and ϵ. The values of λ are 0.001, 0.01, 0.1, 1, 10, 100 and 1000. ϵ are 0.1, 0.5, 1, 1.5, 2, 2.5 and 3. Fig. 8 (a), (b) and (c) show the ACC values according to two parameters when p = 0.2, p = 0.4 and p = 0.6 respectively. We can see that the ACC is sensitive when p = 0.2 and p = 0.4, and the ACC is not sensitive when p = 0.6. Therefore, the performance of clustering is important for parameter selection, especially when there is less prior information.

V. Conclusion

This paper proposes a semi-supervised sparse subspace clustering algorithm based on reweighting. The algorithm not only makes the learned optimal similarity matrix have a good block diagonal structure, but also takes full account of the supervision information. On the one hand, it uses supervision information to construct reasonable similarity matrix; on the other hand, it uses supervision information as

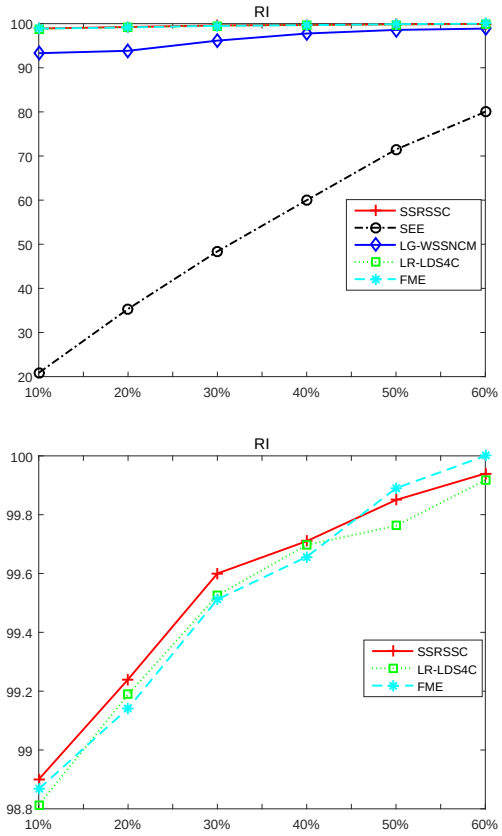

Fig. 5: RI evaluation results of five different semi-supervised clustering algorithms on Coil dataset, the following figure is a zoom in for the above figure.

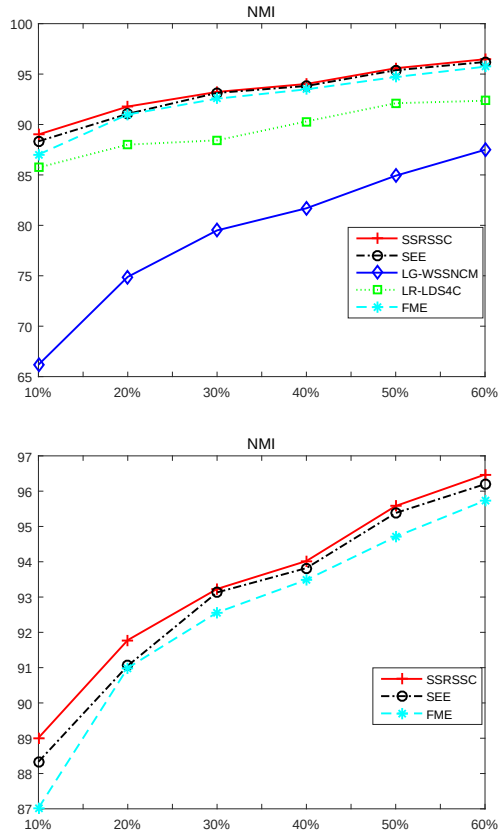

Fig. 6: NMI evaluation results of five different semi-supervised clustering algorithms on USPS dataset, the following figure is a zoom in for the above figure.

label information to guide clustering results. In this paper, the experimental results on benchmark data sets are carried out, and the experimental results further verify the effectiveness of the proposed SSRSSC algorithm. The follow-up research can consider how to ensure its performance is still good while reducing the amount of known label data, and also consider combining with other clustering algorithms.

REFERENCES

[1] A.K. Jain, M.N. Murty and P.J. Flynn, "Data clustering: a review", *ACM Computing Survey*, vol. 31, no.3, pp. 264-323, 1999.

[2] L. Parsons, E. Haque and H. Liu, "Subspace clustering for high dimensional data: a review", *ACM SIGKDD Explorations Newsletter*, vol. 6, no.1, pp.90-105, 2004.

[3] W.Ying, Z. Zhang, T.S. Huang, and J.Y. Lin, "Multibody grouping via orthogonal subspace decomposition", in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2001, pp. II252-II257.

[4] Y. Ma, A.Y. Yang and D.R. Fossum, "Estimation of subspace arrangements with applications in modeling and segmenting mixed data", *SIAM Review*, vol. 50, no.3, pp. 413-458, 2008. DOI: 10.1137/060655523.

[5] S.R. Rao, A.Y. Yang, S.S. Sastry and Y.W. Ma, "Robust algebraic segmentation of mixed rigid- body and planar motions from two views", *International Journal of Computer Vision*, vol. 88, no.3, pp. 425-446, 2010.

[6] J. Ho, M. Yang, J. Lim, K. Lee, and D. Kriegman, "Clustering appearances of objects under varying illumination conditions", in *IEEE Conference Computer Vision and Pattern Recognition* 2003, pp. 11C18.

[7] P.S. Bradley and O.L. Mangasarian, "K - plane clustering". *Journal of Global Optimization*, vol.16, no.1, pp. 23-32, 2000.

[8] M.E. Tipping and C.M. Bishop, "Mixtures of probabilistic principal component analyzers", *Neural Computation*, vol. 11, no.2, pp. 443-482., 1999

[9] Y. Ma, H. Derksen, W.Hong and J. Wright, "Segmentation of multivariate mixed data via lossy data coding and compression", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no.9, pp. 1546-1562, 2007. DOI: 10.1109/TPAMI.2007.1085.

[10] U.V. Luxburg, "A tutorial on spectral clustering", *Statistic and Computing*, vol. 17, no.4, pp. 395-416, 2007.

[11] G.L. Chen, G. Lerman, "Spectral curvature clustering (SCC)", *International Journal of Computer Vision*, vol. 81, no.3, pp. 317-330, 2009.

[12] E. Elhamifar and R. Vidal, "Sparse subspace clustering", in *IEEE Conference on Computer Vision and Pattern Recogenition* 2009, pp. 2790-2797.

[13] E. Elhamifar and R. Vidal, "Sparse subspace clustering: algorithm, theory and applications", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no.11, pp.2765-2781, 2013. DOI: 10.1109/TPAMI.2013.57.

[14] D.S. Pham, S. Budhaditya, D.Phung and S. Venkatesh, "Reweighted sparse subspace clustering via exploitation of constraints", in *IEEE Conference on Computer Vision and Pattern Recognition* 2012, pp. 550-557.

[15] J. Xu, K. Xu, K. Chen and J. Ruan, "Reweighted sparse subspace clustering", *Computer Vision and Image Understanding*, vol. 138, pp. 25-37, 2015.

[16] C. G. Li and R. Vidal, "Structured sparae clustering: a unified optimization framework", in *IEEE Conference on Computer Vision and Image Recognition* 2015, pp. 277-286.

[17] C. You, D. Robinson and R. Vidal, "Scalable sparse subspace clustering by orthogonal matching pursuit", in *IEEE Conference on Computer Vision and Pattern Recognition* 2016, pp. 3918-3927.

[18] J. Li, Y. Kong and Y. Fu, "Sparse subspace clustering by learning approximation $l0$ codes", *AIAA Conference on Artificial Intelligence*, 2017, pp. 2189-2195.

[19] J. Sun, L. Zhang, B. Du and Y.M. Lai, "Band selection using improved sparse subspace clustering for hyperspectral imagery classifycation", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no.6, pp. 2784-2797, 2015.

[20] S. Basu, A.Banerjee and R.J. Mooney, "A probabilistic framework for semi-supervised clustering", in *ACM SICKDD International Conference on Knowledge Discovery and Data Mining* 2004, pp. 59-68.

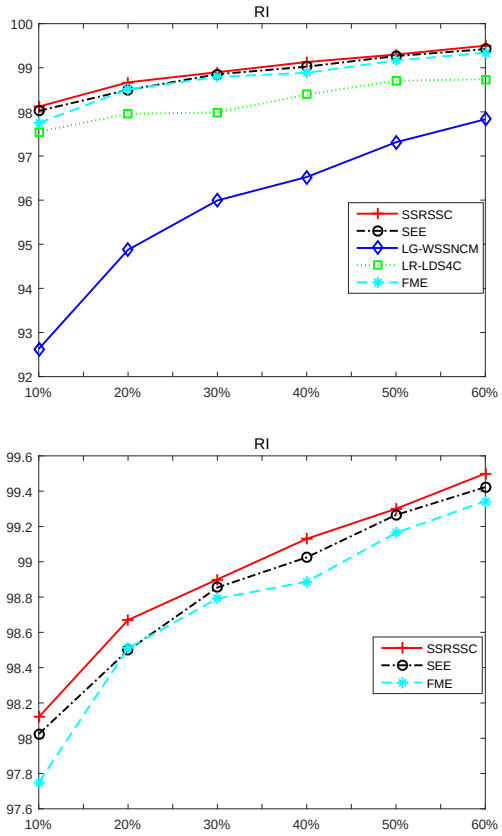

Fig. 7: RI evaluation results of five different semi-supervised clustering algorithms on USPS dataset, the following figure is a zoom in for the above figure.

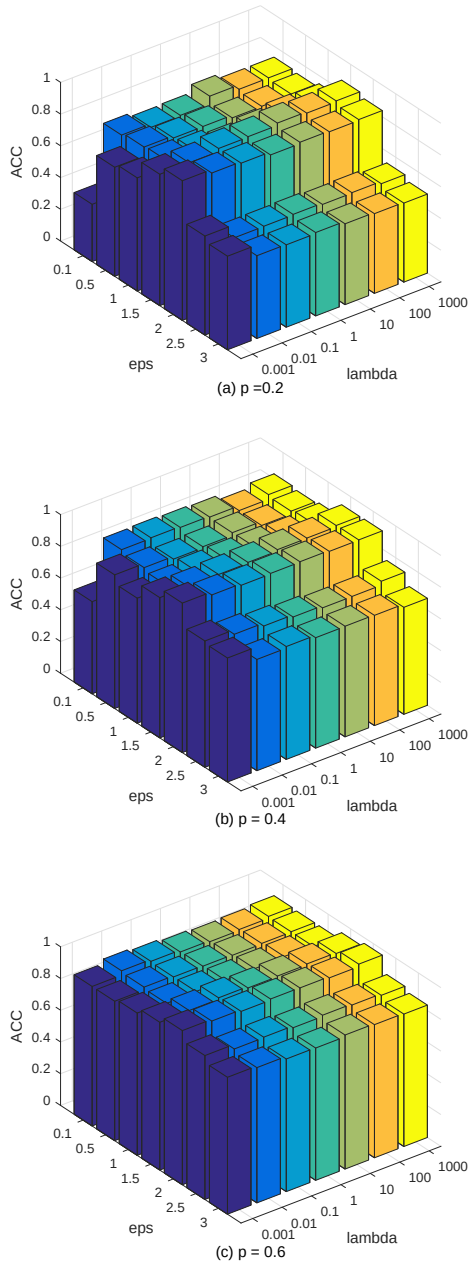

Fig. 8: ACC evaluation results of the given SSRSSC algorithm on Isolet dataset w.r.t two parameters λ and ϵ for different prior information. (a) p=0.2, (b) p=0.4, (c) p=0.6

[21] F. Nie, D. Xu, I. W. Tsang, and C. Zhang, "Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction", *IEEE Transactions on Image Processing*, vol. 19, no. 7, pp. 1921-1932, 2010. DOI: 10.1109/TIP.2010.2044958.

[22] S. Qiu, F. Nie, X. Xu, C. Qing and D. Xu, "Accelerating Flexible Manifold Embedding for Scalable Semi-Supervised Learning". *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 9, pp. 2786-2795, 2019. DOI: 10.1109/TCSVT.2018.2869875.

[23] F. Nie, H. Wang, H. Huang and C. Ding, "Adaptive loss minimization for semi-supervised elastic embedding", in *Proceedings of the 23th International Joint Conference on Artificial Intelligence* 2013, pp.1565-1571. https://www.ijcai.org/Proceedings/13/Papers/233.pdf.

[24] H. Zhu, Y. Ma, "Semi-supervised sparse subspace clustering based on label discrimination and local linear reinforcement", *Application Research of Computers*, vol. 38, no. 10, pp.3014-3018,3034, 2021. DOI: 10.19734/j.issn.1001-3695.2021.03.0044.

[25] D. Zhang, Y. Ma, H. Zhu and F. Smarandache, "A label-guided weighted semi-supervised neutrosophic clustering algorithm", *Journal of Intelligent & Fuzzy Systems*, vol. 43, no. 5, pp. 5661-5672, 2022. DOI: 10.3233/JIFS-212812.

[26] L.S. Zhuang, Z.H. Zhou, S.H. Gao, J. Yin, Z. Lin and Y. Ma, "Label information guided graph construction for semi-supervised learning", *IEEE Transaction on Image Processing*, , vol. 26, no. 9, pp. 4182-4192, 2017.

[27] F. Nie, H. Huang, X. Cai and C. Ding, "Efficient and robust feature selection via joint $L_{2,1}$ norms minimization", in *Proceedings of the 23rd International Conference on Neural Information Processing Systems* 2010, pp. 1813-1821.

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

[29] Z. Lin, R. Liu and Z. Su. "Linearized alternating direction method with adaptive penalty for low rank representation", in *Conference and Workshop on Neural Information Processing Systems* 2011, pp. 612-620.

[30] D. Zhou, T. Bousquet T. Lal, J. Weston and B.S. Olkopf, "Learning with local and global consistency", *Conference and Workshop on Neural Information Processing Systems*, vol. 16, no.3, pp. 595-602, 2003.

[31] G.C. Liu, Z.C. Lin and Y. Yu, "Robust subspace segmentation by low-rank sepresentation", in *International Conference on Machine Learning* 2010, pp. 1172-1180.

[32] J. Wright, A. J. Yang, A. Ganesh, A. Wagner and M.Yi, "Robust face recognition via sparse representation", *IEEE Trans on Pattern Analysis and Machine Intelligence*, vol. 55, no.5, pp. 210-227, 2009.

Qiaoyan Li received her Bachelor Degree on Computational Mathematics from Northwest University in 2000, and Master degree on Applied Mathematics from Xi'an Polytechnic University in 2007. She is currently an associate professor of School of Science, Xi'an Polytechnic University. Her main research interests include statistic learning, fuzzy logic and neutrosophic set theory.

Xue Zhao is currently a postgraduate student of School of Science, Xi'an Polytechnic University, China. Her main research interests include machine learning and computational mathematics.

Hengdong Zhu received his Bachelor Degree on Information and Computing Science from Xi'an Polytechnic University in 2018, and master Degree Computational Mathematics on Xi'an Polytechnic University in 2021. His main research interests include machine learning and computational mathematics.