Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Capped l_1-Norm Sparse Representation Method for Graph Clustering

MULIN CHEN[1], QI WANG[1], SHANGDONG CHEN[2], AND XUELONG LI[1]

[1]School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, P.R. China (e-mail: chenmulin001@gmail.com, crabwq@gmail.com, li@nwpu.edu.cn)

[2]School of Information Science and Technology, Northwest University, Xi'an 710127, China. (e-mail: 201810234@stumail.nwu.edu.cn)

Corresponding author: Xuelong Li (e-mail: li@nwpu.edu.cn).

ABSTRACT As one of the most popular clustering techniques, graph clustering has attracted many researchers in the field of machine learning and data mining. Generally speaking, graph clustering partitions the data points into different categories according to their pairwise similarities. Therefore, the clustering performance is largely determined by the quality of the similarity graph. The similarity graph is usually constructed based on the data points' distances. However, the data structure may be corrupted by outliers. To deal with the outliers, we propose a Capped l_1-norm Sparse Representation method (CSR) in this paper. The main contribution of this work are threefold: (1) a similarity graph with clear cluster structure is learned by employing sparse representation with proper constraints; (2) the capped l_1-norm loss is utilized to remove the outliers, which ensures the graph quality; (3) an iterative algorithm is developed to optimize the proposed non-convex problem. Extensive experiments on real-world datasets show the superiority of the proposed method over the state-of-the-arts, and demonstrate its robustness to the outliers.

INDEX TERMS Graph clustering, graph learning, capped norm, sparse representation.

I. INTRODUCTION

CLUSTERING is an important research area in machine learning and data mining. Given the data points, clustering aims to divide them into clusters, so that the highly correlated points are grouped together. To achieve this goal, a great number of algorithms are proposed in the past decades, such as k-means [1], hierarchical clustering [2], [3], affinity propagation [4], matrix factorization [5]–[7], multiview clustering [8], [9] and graph clustering [10]–[12]. Among them, graph clustering has become one of the most widely used approaches due to the exploration of data manifold, and achieved good performance in a variety of real-world tasks, such as image segmentation [12], crowd group detection [13], and biological sequence data clustering [14].

Graph clustering methods perform clustering based on the similarity graph, which reflects the relationship between data points. According to whether the graph is fixed, existing algorithms can be roughly classified into predefined graph-based approaches and adaptive graph-based approaches. The first kind of methods, such as Ratio Cut [11] and Normalized Cut [12], perform graph construction and clustering in two separate steps. Once the similarity graph is built, it does not change during the optimization stage. However, the predefined graph may be not suitable for the specific clustering task, because it is difficult to select the appropriate kernel [15]. In order to mitigate these problems, some researchers propose to update the graph adaptively during clustering. These adaptive graph-based methods [13], [16]–[21] combine graph learning into the clustering problem, and learn the data points' similarities by investigating the local structure. Nevertheless, the outliers in the input data may corrupt the data structure and affect the quality of the obtained graph. Consequently, it is essential to tackle the outliers.

In recent years, many efforts have been spent on improving the robustness of clustering algorithms. Since the widely used quadric loss function squares the residual error of each data point, the objective value tends to be dominated by outliers. Accordingly, some approaches employ robust loss functions for residual calculation, such as l_1-norm loss [18] and $l_{2,1}$-norm loss [22], [23]. These methods reduce the influence of outliers to some extent. However, the extreme outliers may still affect the results because the aforementioned loss functions can not remove the outliers, but just weakens their effects.

To deal with the above problems, this paper puts forward a Capped l_1-norm Sparse Representation method (CSR) for graph clustering. The proposed model learns the optimal

graph for clustering by integrating sparse representation and capped l_1-norm loss function. The contributions made in this study are summarized as follows:

- Sparse representation is used to learn the similarity graph adaptively. With the suggested constraints, the optimal graph is obtained, which indicates the explicit cluster structure.
- A capped l_1-norm loss-based objective function is designed, which removes the effects of outliers by truncating their residuals. In this way, both the graph quality and the robustness are improved.
- The proposed non-convex problem is solved with an efficient algorithm, and the final result is achieved without any post-processing.

The remaining part of this paper is organized as follows. Section II reviews the related works on graph clustering. Section III introduces the proposed method, and describes the corresponding optimization algorithm. Section IV gives the experimental results on real-world datasets. Section V concludes this paper.

II. RELATED WORK

In this section, the previous works on graph clustering are briefly reviewed.

To cluster the data points with complicated structures, it is essential to capture the local data relationship. Spectral clustering [10] is a representative graph clustering technique. It builds a weighted graph according to the points' local distance, and seeks the clustering consistency between neighbors by utilizing the spectrum of the graph. Chan et al. [11], Hagen et al. [24] and Shi et al. [12] designed different criteria to segment the graph, such that the within-class similarity is maximized and the between-class similarity is minimized. Zelnik-Manor and Perona [25] put forward the self-tuning spectral clustering method to determine the scale of the similarity graph automatically. To partition the points with small distances into the same group, some works [6], [7] introduced the graph regularization term on the formulation of matrix factorization. These methods have shown dominant performance in the clustering literature, however, all of them rely on the predefined similarity graph. It is difficult to select the appropriate graph construction approach for various tasks. Once the graph is built with low quality, these methods are unlikely to get the correct clustering performance.

To alleviate the dependence on the predefined graph, some researchers perform graph learning during the graph-theoretic optimization procedure. Huang et al. [20] and Kang et al. [19] transformed the formulation of sparse representation, and utilized it to learn the data similarity. Nie et al. [17] imposed the rank constraint to learn a similarity graph with clear cluster structure. Wang et al. [13] utilized the manifold ranking technique for graph learning. The above methods adjust the similarity graph during adaptively, so their results are independent on the graph construction strategy. However, the graph quality may be still influenced by the outliers since the graph is updated according to the points' local distances. Some methods [18], [22], [23] employed l_1-norm and $l_{2,1}$-norm to improve the robustness, but the effects of the outliers can not be removed thoroughly.

To reduce the above problems, we proposed a new graph clustering method in this paper. Compared with the previous works, the proposed model has three major advantages: (1) it does not rely on the input graph. The graph structure is learned automatically by sparse representation, which captures the local relationship between data points; (2) it does not need the post-processing procedure. The desired graph indicates the clear cluster structure, so clustering is accomplished once the optimization is over; (3) it is robust to the extreme outliers. The capped l_1-norm removes the affect of a point once it is distinguished as a outlier, so the robustness is ensured.

III. CAPPED L_1-NORM SPARSE REPRESENTATION

In this section, we first revisit the sparse representation method as the preliminary. Then, the proposed Capped l_1-norm Sparse Representation method (CSR) is introduced, and the optimization algorithm is developed.

Throughout this paper, the notations are defined as follows. For a matrix $\mathbf{M}$, its transpose is denoted by $\mathbf{M}^T$. The trace of $\mathbf{M}$ is denoted by $Tr(\mathbf{M})$. The Frobenius-norm of $\mathbf{M}$ is denoted by $||\mathbf{M}||_F$, and the $l_{2,1}$-norm is denoted by $||\mathbf{M}||_{2,1}$. For a vector m, its i-th element is denoted by m_i, and the l_p-norm of m is denoted by $||m||_p$ $(p = 0, 1, 2)$. $\mathbf{1}$ is a column vector with all the elements as 1, and $\mathbf{I}$ is the identity matrix.

A. PRELIMINARIES

Suppose we have n data points $\mathbf{X} = [x_1, ..., x_n] \in \mathbb{R}^{d\times n}$, and each point $x_i \in \mathbb{R}^{d\times 1}$ is a d dimensional vector. Given a new observation $y \in \mathbb{R}^{d\times 1}$, spare representation methods [26], [27] try to represent y with $\mathbf{X}\beta$, where $\beta \in \mathbb{R}^{n\times 1}$ is a sparse coefficient. The objective is formulated as

$$\min_{\beta} ||y - \mathbf{X}\beta||_2^2 + \lambda||\beta||_0, \tag{1}$$

where the first term minimizes the reconstruction error and the second term enforces the coefficient vector to be sparse. However, it is difficult to solve the l_0-norm problem. According to Donoho [28], Eq. (1) can be relaxed as

$$\min_{\beta} ||y - \mathbf{X}\beta||_2^2 + \lambda||\beta||_1, \tag{2}$$

which is easier to solve. With the sparse vector β, the discriminative information is maintained and the computational cost is saved.

Since sparse representation does not need to specify the analysis scale, Huang et al. [20] proposed to extend it into the graph clustering literature, and designed the following objective function:

$$\min_{\alpha_i \geq 0} \sum_{i=1}^{n} (||x_i - \mathbf{X}_{-i}\alpha_i||_2^2 + \lambda||\alpha_i||_1), \tag{3}$$

where $\mathbf{X}_{-i}$ is the dataset without the i-th point and $\alpha_i \in \mathbb{R}^{(n-1)\times 1}$ is the sparse coefficient . The above formulation

assumes that x_i can be reconstructed by the combination of the other points. Intuitively, if α_{ij} is large, x_j contributes the most to the reconstruction of x_i, which implies that the similarity between x_i and x_j should be large. Therefore the $\alpha_i \in \mathbb{R}^{(n-1)\times 1}$ can be also considered as the similarity vector, and the similarity graph is obtained by minimizing the reconstruction error of each point. Furthermore, Huang et al. [20] removed the sparse term $\lambda||\alpha_i||_1$ by imposing an additional constraint $\alpha_i^T\mathbf{1}=1$, which yields to a closed-form objective function:

$$\min_{\alpha_i\geq 0,\alpha_i^T\mathbf{1}=1}\sum_{i=1}^{n}||x_i-\mathbf{X}_{-i}\alpha_i||_2^2. \tag{4}$$

B. PROBLEM FORMULATION

In this part, the proposed Capped l_1-norm Sparse Representation method (CSR) is presented. Denoting the desired similarity graph as $\mathbf{S}\in\mathbb{R}^{n\times n}$, we transform problem (5) into

$$\min_{s_i\geq 0,s_i^T\mathbf{1}=1}\sum_{i=1}^{n}||x_i-\mathbf{X}s_i||_2^2+\gamma||\mathbf{S}||_F^2. \tag{5}$$

where $s_i\in\mathbb{R}^{n\times 1}$ is the i-th column of $\mathbf{S}$, and γ is a regularization parameter. The regularization term $||\mathbf{S}||_F^2$ avoids the trivial solution, where the optimal $\mathbf{S}$ is the identity matrix. As discussed in Section I, the quadric loss function in the first term of problem (5) is sensitive to the data outliers. To penalize the outliers, we improve the sparse representation method by introducing the capped l_1-norm loss, which leading to the following problem:

$$\min_{s_i\geq 0,s_i^T\mathbf{1}=1}\sum_{i=1}^{n}\min(||x_i-\mathbf{X}s_i||_2,\varepsilon)+\gamma||\mathbf{S}||_F^2, \tag{6}$$

where ε is a threshold. In problem (6), x_i is considered as outlier if its reconstruction error $||x_i-\mathbf{X}s_i||_2$ is larger than ε, which means it cannot be well approximated by the others. Then its residual becomes a constant. In this way, the effects of extreme outliers are removed. For the points with small reconstruction errors, the first term of problem (6) becomes the standard $l_{2,1}$-norm. Therefore, the proposed objective function is more robust than the $l_{2,1}$-norm formulations theoretically.

By solving problem (6), a sparse similarity graph is learned. However, the obtained graph $\mathbf{S}$ does not have the clear cluster structure, which means that we have to perform post-processing (k-means, spectral clustering) to get the clustering result. Denoting the Laplacian matrix of $\mathbf{S}$ as $\mathbf{L_S}$, Mohar et al. [29] have proved that the number of zero eigenvalues of $\mathbf{L_S}$ equals to the number of connected components of $\mathbf{S}$. Ideally, supposing the desired cluster number is c, the optimal similarity graph $\mathbf{S}$ should contain exactly c connected components, where the points from the same class are connected into one component. Denoting the i-th smallest eigenvalue of $\mathbf{L_S}$ as $\sigma_i(\mathbf{L_S})$, $\sigma_i(\mathbf{L_S})$ is non-negative because $\mathbf{L_S}$ is positive semi-definite. Then we enforce $\mathbf{S}$ to contain c connected components by solving

$$\min\sum_{i=1}^{c}\sigma_i(\mathbf{L_S}). \tag{7}$$

According to Nie et al. [30], problem (7) is equivalent to the following problem

$$\min_{\mathbf{F}\in^{n\times c},\mathbf{F}^T\mathbf{F}=\mathbf{I}}Tr(\mathbf{F}^T\mathbf{L_S}\mathbf{F}). \tag{8}$$

Combining Eq. (6) and (8), we have the objective function of the proposed CSR method

$$\begin{aligned}&\min_{\mathbf{S},\mathbf{F}}\sum_{i=1}^{n}\min(||x_i-\mathbf{X}s_i||_2,\varepsilon)+\gamma||\mathbf{S}||_F^2+\lambda Tr(\mathbf{F}^T\mathbf{L_S}\mathbf{F}),\\&s.t.s_i\geq 0,s_i^T\mathbf{1}=1,\mathbf{F}\in^{n\times c},\mathbf{F}^T\mathbf{F}=\mathbf{I},\end{aligned} \tag{9}$$

where λ is a large enough parameter to enforce $Tr(\mathbf{F}^T\mathbf{L_S}\mathbf{F})$ to be zero. With the above formulation, the proposed method is able to learn the optimal similarity graph $\mathbf{S}$. Since $\mathbf{S}$ consists of c connected components, it can be considered as the cluster indicator matrix with each component corresponding to a cluster.

In problem (9), each point is approximated by the linear combination of the others. For points x_i and x_j, their similarity s_{ij} becomes larger if x_j contributes more to the reconstruction of x_i. In this way, the similarity is adjusted according to the points' local relationship, and the graph quality is improved.

C. OPTIMIZATION ALGORITHM

Here the optimization strategy for solving problem (9) is designed. The similarity graph $\mathbf{S}$ is initialized with an efficient approach [17]. Problem (9) involves two variables to be optimized, so we fix one and update the other one iteratively.

Fix S update F

When $\mathbf{S}$ is fixed, the objective function becomes problem (8). With the orthogonal constraint, the minimum value of $Tr(\mathbf{F}^T\mathbf{L_S}\mathbf{F})$ is $\prod_{i=1}^{c}\sigma_i(\mathbf{L_S})$. Thus, the optimal $\mathbf{F}$ is formed with the c eigenvectors associated with the c smallest eigenvalues of $\mathbf{L_S}$.

Fix F update S

When $\mathbf{F}$ is fixed, denoting the i-th row of $\mathbf{F}$ as f_i, we have

$$Tr(\mathbf{F}^T\mathbf{L_S}\mathbf{F})=\frac{1}{2}\sum_{i,j}||f_i-f_j||_2^2s_{ij}, \tag{10}$$

where s_{ij} is the j-th element in s_i. Then problem (9) becomes

$$\begin{aligned}&\min_{\mathbf{S}}\sum_{i=1}^{n}\min(||x_i-\mathbf{X}s_i||_2,\varepsilon)+\gamma\sum_{i=1}^{n}s_i^Ts_i+\\&\frac{1}{2}\lambda\sum_{i=1}^{n}||f_i-f_j||_2^2s_{ij},\\&s.t.s_i\geq 0,s_i^T\mathbf{1}=1,\end{aligned} \tag{11}$$

which is non-convex. We propose to solve it with the re-weighted algorithm [30]. It is easy to devise that the derivative of problem (11) can be approximated by the derivative of the following problem

$$\begin{aligned} \min_{\mathbf{S}} & \sum_{i=1}^{n} d_i ||x_i - \mathbf{X}s_i||_2^2 + \gamma \sum_{i=1}^{n} s_i^T s_i + \\ & \frac{1}{2}\lambda \sum_{i=1}^{n} ||f_i - f_j||_2^2 s_{ij}, \\ s.t. & s_i \geq 0, s_i^T \mathbf{1} = 1, \end{aligned} \tag{12}$$

where

$$d_i = \begin{cases} \frac{1}{2||x_i - \mathbf{X}\tilde{s}_i||_2}, & if\ ||x_i - \mathbf{X}\tilde{s}_i||_2 < \varepsilon \\ 0, & else \end{cases}, \tag{13}$$

and $\tilde{s}_i$ is the current solution. Nie et al. [30] have proved that problem (12) will finally converge to the optimal solution to problem (11), so the desired $\mathbf{S}$ can be learned by solving problem (12) iteratively. Fixing d_i with $\tilde{s}_i$, for each point x_i, we need to solve the following problem

$$\min_{s_i \geq 0, s_i^T \mathbf{1}=1} d_i ||x_i - \mathbf{X}s_i||_2^2 + \gamma s_i^T s_i + \frac{1}{2}\lambda p_i^T s_i, \tag{14}$$

where $p_i \in \mathbb{R}^{n\times 1}$ is a vector with its j-th element equal to $||f_i - f_j||_2^2$. When $d_i > 0$, by removing the irrelevant terms, we get

$$\min_{s_i \geq 0, s_i^T \mathbf{1}=1} s_i^T (\mathbf{X}^T\mathbf{X} + \gamma \mathbf{I}) s_i + (\frac{\lambda}{2d_i} p_i^T - 2x_i^T \mathbf{X}) s_i, \tag{15}$$

and when $d_i = 0$, we have

$$\min_{s_i \geq 0, s_i^T \mathbf{1}=1} s_i^T s_i + \frac{\lambda}{2d_i} p_i^T s_i, . \tag{16}$$

Both problem (15) and (16) can be efficiently solved by the Augmented Largrange Method (ALM) [31].

By updating $\mathbf{F}$ and $\mathbf{S}$ iteratively, the optimal similarity graph with c connected components is learned. The detailed optimization algorithm is shown in Algorithm 1. Note that, in the implementation, the parameter λ is tuned in a heuristic strategy according to the number of connected components in $\mathbf{S}$, as described in Algorithm 1.

Algorithm 1 Optimization algorithm of CSR

Input: Data matrix $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, parameter γ and λ, cluster number c.

Initialize similarity graph $\mathbf{S}$.
repeat
  Update $\mathbf{F}$ by solving problem (8);
  for each point x_i **do**
    Update d_i with the current $\tilde{s}_i$ by Eq. (13);
    Update s_i by solving problem (15) or (16) according to the value of d_i;
  end for
  if $\mathbf{S}$ has more than c connected components **then**
    $\lambda = \frac{1}{2}\lambda$;
  else if $\mathbf{S}$ has less than c connected components **then**
    $\lambda = 2\lambda$;
  else
    break;
  end if
until Converge

Output: The optimal $\mathbf{F}$ and $\mathbf{S}$.

IV. EXPERIMENTS

In this section, extensive experiments are conducted to validate the effectiveness of the proposed Capped l_1-norm Sparse Representation method (CSR). The widely used clustering accuracy (ACC) and Normalized Mutual Information (NMI) [17] are used as evaluation measurements. All the experiments are implemented in MATLAB R2015b , and run on a Windows 8 machine with 3.20 GHz i5-3470 CPU, 32 GB main memory.

A. EXPERIMENT SETTING

In this work, experiments are preformed on eight real-world benchmarks, which are briefly described as follows.

- Jaffe [32] is a face dataset that contains 213 images from 10 classes. The images are captured with different facial expressions, which are cropped as 26×26 pixels.
- Extended Yale B [33] (shorten as Yale B) includes 2414 cropped face images from 38 classes. We randomly select 10 images from each class, and utilize the selected 380 samples for evaluation. Each image is with 32×32 pixels.
- Umist [34] is also a face dataset, which contains 575 images from 20 subjects. The subjects cover different races, sexes and appearances. Each image is resized into 28×23 pixels.
- USPS [35] is consisted of 9298 digit images, and the handwritten number ranges from 0 to 9. We randomly select 20 images from each number and the image size is 28×28.
- Mnist [36] includes 70000 digit images. There are 10 classes in total. For each class, 50 images are selected. The images are resized into 28×28 pixels.
- Lung [37] comprises 203 gene expression sequences of lung speciments, including 4 kinds of lung tumors and one kind of normal lung samples. Each sequence is represented by 3312 genes.
- Isolet5 and Mfeat-pix are from the UCI Machine Learning Repository [38]. Isolet5 collects the spoken letter from human speakers. Mfeat-pix is constructed with the features of handwritten images.

The detailed information of the datasets is listed in Table 1.

To demonstrate the effectiveness of the proposed CSR method, seven state-of-the-art clustering methods are em-

TABLE 2. ACC of the clustering methods on nine datasets.

Datasets	k-means	RCut	NCut	NMF	CAN	SSR	RASL	CSR
Jaffe	0.7429	0.7437	0.7378	0.6244	0.7042	0.7923	0.8313	**0.9671**
Yale B	0.4662	0.1774	0.5711	0.4553	0.7421	0.7171	0.7258	**0.7500**
Umist	0.4231	0.3918	0.4560	0.3478	0.6939	0.4726	0.5134	**0.7148**
USPS	0.5648	0.2313	0.5543	0.4000	0.4750	0.6077	0.6133	0.68
Mnist	0.4666	0.2402	0.4520	0.4660	0.3320	0.5534	0.5140	**0.5680**
Lung	0.7407	0.6121	0.5421	0.6207	0.7931	0.6305	0.5785	**0.8818**
Isolet5	0.5655	0.3262	0.5705	0.3603	0.4346	0.5640	0.5624	**0.5827**
Mfeat-pix	0.6874	0.4055	0.6675	0.5005	0.7680	0.7847	0.7152	**0.8295**

TABLE 3. NMI of the clustering methods on nine datasets.

Datasets	k-means	RCut	NCut	NMF	CAN	SSR	RASL	CSR
Jaffe	0.8313	0.7979	0.8125	0.6723	0.8512	0.8599	0.8870	**0.9623**
Yale B	0.6470	0.3132	0.7372	0.6407	0.7676	0.7376	0.7567	**0.7901**
Umist	0.6368	0.5579	0.6350	0.4887	0.8536	0.6334	0.7055	**0.8637**
USPS	0.5476	0.2139	0.5545	0.3982	0.4607	0.6325	0.6389	0.6912
Mnist	0.4502	0.2108	0.4556	0.4230	0.3118	0.5245	0.4883	**0.5483**
Lung	0.5454	0.4259	0.4108	0.3757	0.5148	0.4797	0.3926	**0.6970**
Isolet5	0.7226	0.4989	0.7111	0.4701	0.6221	0.7285	0.7306	**0.7410**
Mfeat-pix	0.7021	0.4359	0.6800	0.4711	0.8051	0.7936	0.7837	**0.8469**

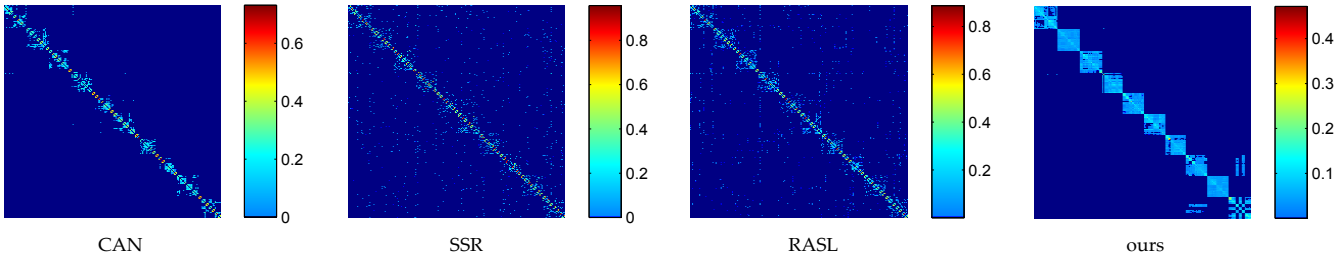


FIGURE 1. The similarity graph learned by (a) CAN, SSR, RASL and the proposed CSR. The cluster structure in the graph of CSR is clear. Best viewd in color.

TABLE 1. The detailed information of the real-world datasets.

Datasets	Samples	Dimensions	Classes
Jaffe	213	676	10
Yale B	380	1024	38
Umist	575	644	20
USPS	200	256	10
Mnist	500	784	10
Lung	213	3312	5
Isolet5	1560	617	26
Mfeat-pix	2000	240	6

ployed for comparison, including k-means, Ratio Cut (RCut) [11], Normalized Cut (NCut) [12], Non-negative Matrix Factorization (NMF) [5], Clustering with Adaptive Neighbors (CAN) [17], Simplex Sparse Representation (SSR) [20] and Robust Adaptive Sparse Learning (RASL) [23]. For the predefined-graph based methods, such as RCut and NCut, the self-tune Gaussian method [25] is used to construct the similarity graph. For CAN, SSR, RASL and the proposed CSR, an efficient method [17] is utilized to initialize the graph, and the neighborhood size is 5. Because k-means, RCut and NCut are sensitive to the initialization, we repeat them for 30 times and report the averaged results. The graphs learned by SSR and RASL need to be processed to get the final results, so we repeat NCut on the learned graphs for 30 times and show the averaged results. For a fair comparison, we let each competitor use its optimal parameters. For the proposed method, we simply set γ and the initial value of λ as 1.

B. PERFORMANCE WITHOUT EXPLICIT OUTLIERS

We first show the clustering results of the eight methods on datasets without explicit outliers. For each dataset, we tune the value of ε to find 2 percentage outliers. As shown in Table 2 and 3, the proposed CSR has the highest ACC and NMI on all the datasets. Especially, CSR outperforms the second best method a lot on Jaffe and Lung. Both k-means and NMF fail to exploit the local manifold, so they can not deal with the data with complex structures. RCut and NCut learn the data relationship with the Gaussian similarity graph and try to achieve the clustering consistency between local neighbors, but the predefined graph may be not suitable for clustering. CAN, SSR and RASL obtain relatively better results, because they integrate graph learning into the clustering procedure. However, the graph quality is not guaranteed since both the l_2-norm loss and $l_{2,1}$-norm loss can not remove the effects of outliers. The proposed CSR optimizes the similarity graph by minimizing the reconstruction error of each point, and

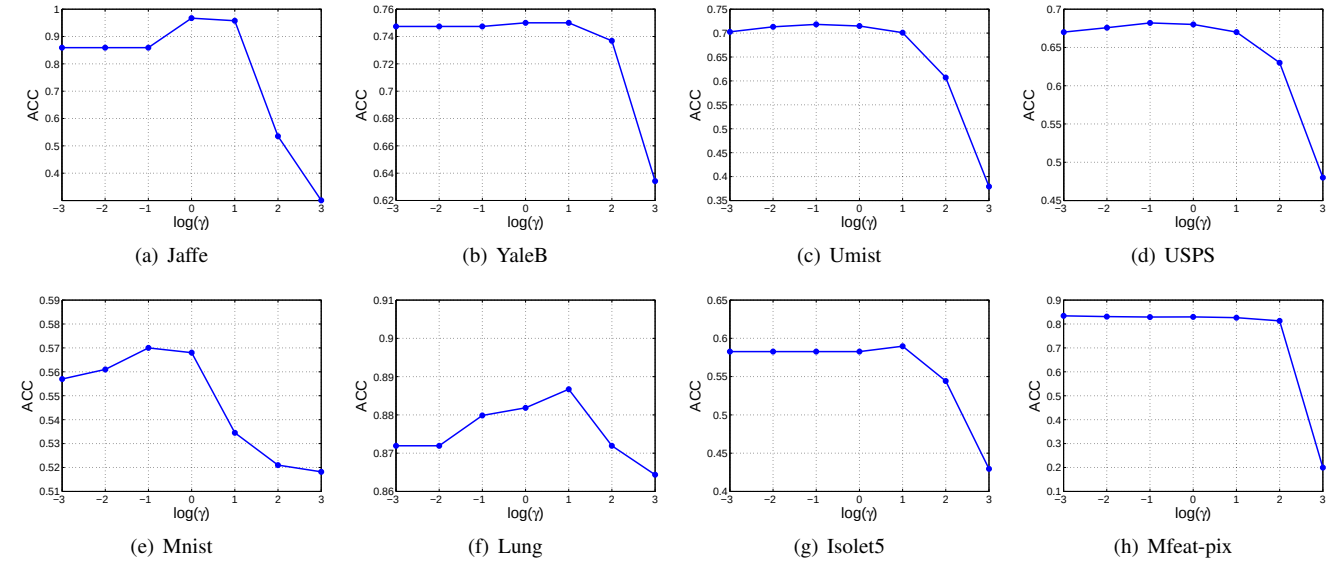


FIGURE 2. **The ACC of CSR curves with respect to the value of γ on eight datasets.**

truncates the residuals of outliers with the capped l_1-norm objective function. Consequently, CSR shows the best performance.

We also visualize the graphs learned by CAN, SSR, RASL and CSR on Jaffe, as shown in Fig. 1. It is manifest that the graph of CSR has the clear cluster structure. While the graphs learned by CAN, SSR and RASL just assign high similarities to the neighbors. Particularly, the graphs obtained by SSR and RASL are with serve noises, which means that many points are incorrectly connected. In the graph of CSR, each connected component indicates a cluster, and only a small portion of the between-class points are connected. Therefore, the proposed method is able to learn a similarity graph with high quality.

In addition, the parameter sensitivity of the proposed method is investigated. The objective function involves two regularization parameters, i.e. γ and λ. As described in Section III-C, λ is self-tuned, so we only need to discuss the effect of γ, which controls the weight of the smooth term in formulation (9). Figure 2 plots the ACC of the proposed algorithm by varying γ from 10^{-3} to 10^3. As shown in the curves, our method achieves stable performance when γ is within the range of $\{10^{-3}, \cdots, 1\}$.

C. PERFORMANCE WITH EXPLICIT OUTLIERS

To verify the robustness of CSR, we show its clustering results on datasets with extreme outliers. We add 30 images, which are collected from different classes of the Caltech101 dataset [39], to the Jaffe dataset, as shown in Figure 3. Each added image is inconsistent with any other one, so they are treated as extreme outliers. The constructed datasets contains 11 classes, where one class is consisted of the outlier images. We evaluate the capability of our method to identify the outlier class and partition the normal images correctly.

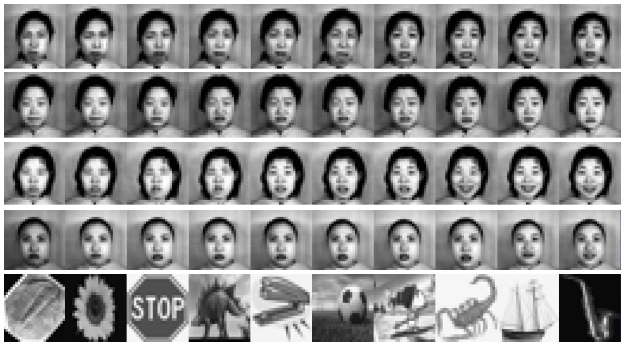


FIGURE 3. **The Jaffe dataset with the outliers from Caltech101. The first four rows are normal face images, and the bottom row visualizes the added outliers. Only some representative images are shown.**

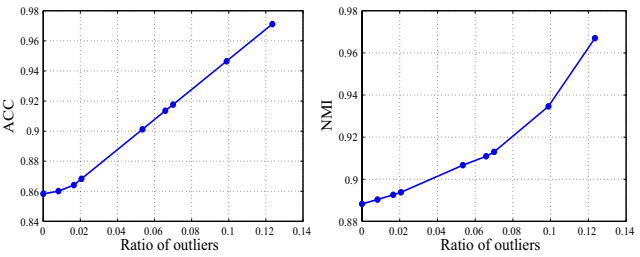
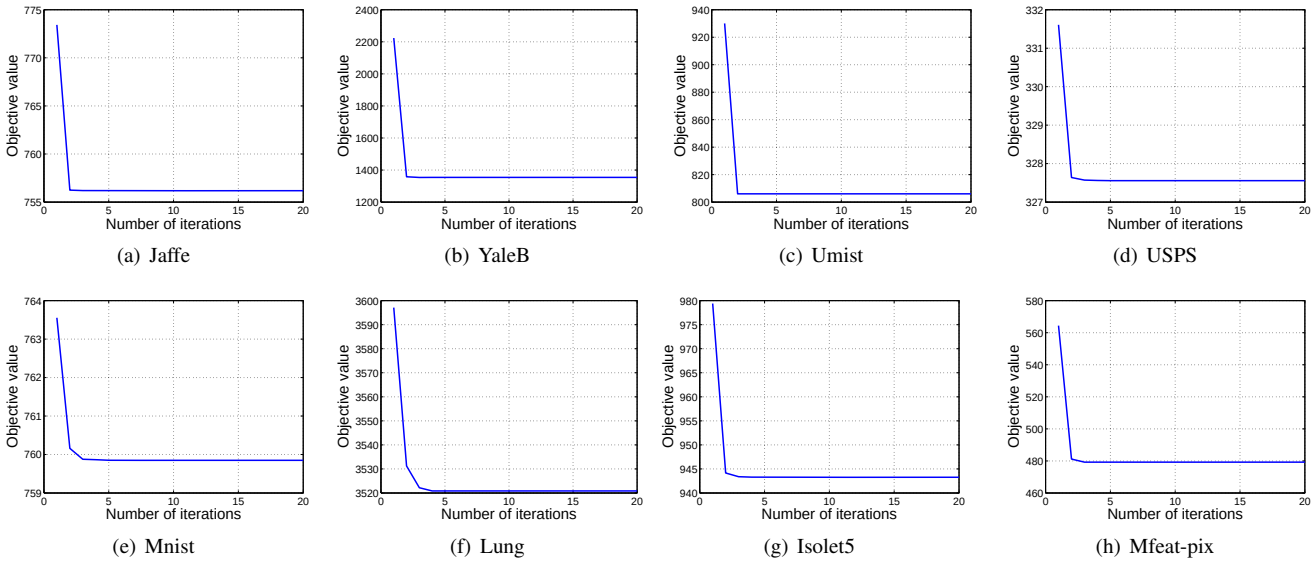


FIGURE 4. **The clustering performance of CSR with respect to the ratio of the outliers on the Jaffe dataset with outliers.**

The ACC and NMI curves with respect to the ratio of outliers are plotted in Figure 4. When the ratio is small, the remained outliers impact the clustering performance. When the ratio becomes larger, the ACC and NMI increase steadily because the extreme outliers are correctly removed. The clustering results of the competitors are exhibited in Table 4. CSR outperforms the other methods to a great extent. Thus,

TABLE 4. Clustering results on the Jaffe dataset with outliers.

Metrics	k-means	RCut	NCut	NMF	CAN	SSR	RASL	CSR
ACC	0.6044	0.7193	0.7519	0.4897	0.3169	0.7815	0.7728	**0.9712**
NMI	0.7349	0.7910	0.8146	0.5725	0.3498	0.8323	0.8250	**0.9670**



(a) Jaffe (b) YaleB (c) Umist (d) USPS

(e) Mnist (f) Lung (g) Isolet5 (h) Mfeat-pix

FIGURE 5. Convergence curves of CSR on eight datasets.

the utilization of the capped l_1-norm loss does improve the robustness to the outliers.

D. CONVERGENCE ANALYSIS

Here we discuss the convergence behavor of the proposed method. The objective function is approximated by problem (12) with the re-weighted method. Nie et al. [30] have proved that the objective value of formulation (9) decreases during the optimization of problem (12). When updating $\mathbf{F}$, the global minimum solution is obtained. When updating $\mathbf{S}$, the ALM algorithm searches a local optimal value. Therefore, the objective value decreases monotonically during optimizing each variable, and finally converges to a local optima. The convergence curves of problem (9) are shown in Figure 5. The objective value converges within 5 iterations on all the dataset, which verifies the efficiency of the proposed optimization algorithm.

V. CONCLUSIONS

In this research, a new graph clustering method termed as Capped l_1-norm Sparse Representation (CSR) has been presented. The proposed model utilizes the capped l_1-norm loss to handle the outliers with large fitting errors, so the robustness is improved. In addition, by employing sparse representation with proper constraints, CSR is able to learn the optimal similarity graph, which indicates the clustering result explicitly. Although the proposed objective function is not convex, it can be readily solved by the suggested optimization algorithm. Experimental results on real-world datasets show the state-of-the-art clustering performance of the proposed method.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China under Grant 2018YFB1107403, National Natural Science Foundation of China under Grant U1864204 and 61773316, State Key Program of National Natural Science Foundation of China under Grant 61632018, Natural Science Foundation of Shaanxi Province under Grant 2018KJXX-024, and Project of Special Zone for National Defense Science and Technology Innovation.

REFERENCES

[1] J. Macqueen, "Some methods for classification and analysis of multivariate observations," in Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281–297.
[2] S. Zafar, A. Bashir, and S. A. Chaudhry, "Mobility-aware hierarchical clustering in mobile wireless sensor networks," IEEE Access, vol. 7, pp. 20 394–20 403, 2019.
[3] F. Rohlf, "Adaptive hierarchical clustering schemes," Systematic Zoology, vol. 19, no. 1, pp. 58–82, 1970.
[4] D. Dueck and B. Frey, "Non-metric affinity propagation for unsupervised image categorization," in International Conference on Computer Vision, pages = 1–8, year = 2007,.
[5] T. Li and C. Ding, "The relationships among various nonnegative matrix factorization methods for clustering," in IEEE International Conference on Data Mining, 2006, pp. 362–371.
[6] J. Huang, F. Nie, H. Huang, and C. Ding, "Robust manifold nonnegative matrix factorization," ACM Transactions on Knowledge Discovery from Data, vol. 8, no. 3, pp. 11:1–11:21, 2013.
[7] D. Cai, X. He, J. Han, and T. Huang, "Graph regularized nonnegative matrix factorization for data representation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 8, pp. 1548–1560, 2011.

[8] Q. Wang, M. Chen, F. Nie, and X. Li, "Detecting coherent groups in crowd scenes by multiview clustering," in IEEE Transactions on Pattern Analysis on Machine Intelligence, 2018.
[9] F. Nie, J. Li, and X. Li, "Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification," in International Joint Conference on Artificial Intelligence, 2016, pp. 1881–1887.
[10] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in Advances in Neural Information Processing System, 2001, pp. 849–856.
[11] P. Chan, M. Schlag, and J. Zien, "Spectral k-way ratio-cut partitioning and clustering," IEEE Transactions on CAD of Integrated Circuits and Systems, vol. 13, no. 9, pp. 1088–1096, 1994.
[12] J. Shi and J. Malik, "Normalized cuts and image segmentation," IEEE Transactions on Pattern Analysis on Machine Intelligence, vol. 22, no. 8, pp. 888–905, 2000.
[13] X. Li, M. Chen, F. Nie, and Q. Wang, "A multiview-based parameter free framework for group detection," in AAAI Conference on Artificial Intelligence, 2017, pp. 4147–4153.
[14] W. Pentney and M. Meila, "Spectral clustering of biological sequence data," in National Conference on Artificial Intelligence, 2005, pp. 845–850.
[15] Y. Wang, X. Liu, Y. Dou, and R. Li, "Multiple kernel clustering framework with improved kernels," in International Joint Conference on Artificial Intelligence, 2017, pp. 2999–3005.
[16] X. Wang, R. Chen, Z. Zeng, C. Hong, and F. Yan, "Robust dimension reduction for clustering with local adaptive learning," IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 3, pp. 657–669, 2019.
[17] F. Nie, X. Wang, and H. Huang, "Clustering and projected clustering with adaptive neighbors," in ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014, pp. 977–986.
[18] F. Nie, X. Wang, M. Jordan, and H. Huang, "The constrained laplacian rank algorithm for graph-based clustering," in AAAI Conference on Artificial Intelligence, 2016, pp. 1969–1976.
[19] Z. Kang, C. Peng, and Q. Cheng, "Twin learning for similarity and clustering: A unified kernel approach," in AAAI Conference on Artificial Intelligence, 2017, pp. 2080–2086.
[20] J. Huang, F. Nie, and H. Huang, "A new simplex sparse learning model to measure data similarity for clustering," in International Joint Conference on Artificial Intelligence, 2015, pp. 3569–3575.
[21] L. Zhang, Q. Zhang, B. Du, D. Tao, and J. You, "Robust manifold matrix factorization for joint clustering and feature extraction," in AAAI Conference on Artificial Intelligence, 2017, pp. 1662–1668.
[22] L. Zhang, Q. Zhang, B. Du, J. You, and D. Tao, "Adaptive manifold regularized matrix factorization for data clustering," in International Joint Conference on Artificial Intelligence,, 2017, pp. 3399–3405.
[23] M. Chen, Q. Wang, and X. Li, "Robust adaptive sparse learning method for graph clustering," in International Conference on Image Processing, 2018, pp. 1618–1622.
[24] L. Hagen and A. Kahng, "New spectral methods for ratio cut partitioning and clustering," IEEE Transactions on CAD of Integrated Circuits and Systems, vol. 11, no. 9, pp. 1074–1085, 1992.
[25] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in Advances in Neural Information Processing Systems, 2004, pp. 1601–1608.
[26] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 2, pp. 210–227, 2009.
[27] Y. Li, S. Ding, B. Tan, H. Zhao, and Z. Li, "Sparse representation based on the analysis model with optimization on the stiefel manifold," IEEE Access, vol. 7, pp. 8385–8397, 2019.
[28] D. Donoho, "For most large under-determined systems of linear equations the minimal l_1-norm solution is also the sparsest solution," in Communication on Pure and Applied Mathematics, 2004, pp. 797–829.
[29] B. Mohar, Y. Alavi, G. Chartrand, O. R. Oellermann, and A. J. Schwenk, "The laplacian spectrum of graphs," in Graph Theory, Combinatorics, and Applications, 2001, pp. 871–898.
[30] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint l_{21}-norms minimization," in Advances in Neural Information Processing Systems, 2010, pp. 1813–1821.
[31] F. Nie, H. Wang, H. Huang, and C. Ding, "Joint schatten p -norm and ℓ_p -norm robust matrix completion for missing value recovery," Knowledge and Information Systems, vol. 42, no. 3, pp. 525–544, 2015.
[32] M. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, no. 12, pp. 1357–1362, 1999.
[33] A. Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 6, pp. 643–660, 2001.
[34] D. Graham and N. Allinson, "Characterising virtual eigensignatures for general purpose face recognition," Face Recognition Form Theory to Applications, vol. 163, no. 2, pp. 446–456, 1998.
[35] J. Hull, "A database for handwritten text recognition research," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 16, no. 5, pp. 550–554, 1994.
[36] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.
[37] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, and J. P. Richie, "Gene expression correlates of clinical prostate cancer behavior," Cancer Cell, vol. 1, no. 2, p. 203, 2002.
[38] M. Lichman, "UCI machine learning repository," 2013.
[39] F. Li, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," Computer Vision and Image Understanding, vol. 106, no. 1, pp. 59–70, 2007.

MULIN CHEN received the B.E. degree in software engineering and the M.E. degree in computer application technology from Northwestern Polytechnical University, Xi'an, China, in 2014 and 2016 respectively. He is currently pursuing the Ph.D. degree with the Center for Optical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. His current research interests include computer vision and machine learning.

QI WANG (M'15-SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Computer Science, with the Unmanned System Research Institute, and with the Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.

SHANGDONG CHEN received his master degree from Air Force Engineering University, Xi'an, China, in 2008. He is currently a PhD student at School of Information Science and Technology, Northwest University, Xi'an China. His research interests are in the broad areas of signal processing, machine learning, and internet of things.

PLACE PHOTO HERE

XUELONG LI (M'02-SM'07-F'12) is currently a Full Professor with the School of Computer Science and with the Center for OPTical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, China.