Learning with Adaptive Neighbors for Image Clustering

Yang Liu[1], **Quanxue Gao**[1*], **Zhaohua Yang**[2†], **Shujian Wang**[1]
[1] State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an, China
[2] Beihang University, Beijing, China
xidianliuyang@gmail.com, qxgao@xidian.edu.cn, yangzh@buaa.edu.cn, wangsj079@163.com

Abstract

Due to the importance and efficiency of learning complex structures hidden in data, graph-based methods have been widely studied and get successful in unsupervised learning. Generally, most existing graph-based clustering methods require post-processing on the original data graph to extract the clustering indicators. However, there are two drawbacks with these methods: (1) the cluster structures are not explicit in the clustering results; (2) the final clustering performance is sensitive to the construction of the original data graph. To solve these problems, in this paper, a novel learning model is proposed to learn a graph based on the given data graph such that the new obtained optimal graph is more suitable for the clustering task. We also propose an efficient algorithm to solve the model. Extensive experimental results illustrate that the proposed model outperforms other state-of-the-art clustering algorithms.

1 Introduction

Clustering, which aims to partition the data points from different groups such that the samples in the same group have high similarity to each other, has been playing an important role in data mining applications. Lots of clustering algorithms have been proposed in the past decades, such as K-means clustering [MacQueen and others, 1967; Gao *et al.*, 2018], hierarchical clustering [Johnson, 1967], support vector clustering [Ben-Hur *et al.*, 2001; Liu *et al.*, 2018], maximum margin clustering [Xu *et al.*, 2005; Wang *et al.*, 2017], multi-view clustering [Cai *et al.*, 2013; Wang *et al.*, 2013]. As the manifold information and data graph are utilized in the clustering model, the graph-based clustering method often outperforms K-means method, especially for a small number of clusters.

Some clustering methods based on graphical representations of the relationships among data have achieved the state-of-the-art performance, including ratio cut [Hagen and Kahng, 1992], spectral clustering [Ng *et al.*, 2002] and normalized cut [Shi and Malik, 2000]. [Ding *et al.*, 2005] gave a discussion of the relations among these graph-based clustering methods. To enhance the learning results, graph-based learning methods often require pre-processing on the affinity matrix. The doubly stochastic matrix (also called bistochastic matrix) was utilized to normalize the affinity matrix and showed promising clustering results [Zass and Shashua, 2007]. For example, [Nie *et al.*, 2011] attempted to cluster data in a high dimension. [Nie *et al.*, 2014] constructed the similarity matrix by adaptively selecting reliable neighbors of data. Additionally, sparse subspace clustering (SSC) was proposed in [Elhamifar and Vidal, 2013] to explore a sparse representation corresponding to the data from the same subspace. After obtaining the representation of the subspace, the spectral clustering can be performed on such new representation. The low-rank subspace segmentation was proposed in [Liu *et al.*, 2013] to find the subspace structure with a low-rank representation.

Although all above clustering techniques have achieved state-of-the-art performance, they still exist some disadvantages. (1) Such methods tend to conduct the similarity matrix from original data but rarely modify it. The real world datasets often contain outliers and noise that result in the unreliable and inaccurate matrix which will impair the finally performance. (2) These methods involve a two-stage process in which a graph is formed from the data points, and then various optimization procedures are invoked on this fixed input data graph [Nie *et al.*, 2016], which result in the final clustering structures are not represented explicitly in the data graph. (3) They are sensitive to the particular graph construction methods because the final clustering results are dependent on the quality of the input data graph.

In this paper a novel graph-based clustering model is proposed to learn a graph with exactly c connected components (where c is the number of clusters). Instead of fixing the original data graph associated to the affinity matrix, the proposed model tries to learn a new similarity block diagonal matrix which has c connected components. Then, the obtained similarity matrix can be directly used for the final clustering task, rather than requiring post-processing to extract the clustering indicators. To guarantee the existence of exactly c connected components, in our model, a rank constraint on the Laplacian graph of the similarity matrix is imposed. A non-greedy iterative algorithm is presented to efficiently solve our model

*Corresponding author: Q. Gao. (qxgao@xidian.edu.cn)
†Corresponding author: Z. Yang. (yangzh@buaa.edu.cn)

and the convergence of the proposed optimization method is also proved. Experimental results on real-world benchmark datasets illustrate that our proposed method outperforms other related methods in most cases.

Notation: Throughout the paper, all the matrices are written as uppercase. For a matrix $\mathbf{M} \in \mathbf{R}^{d\times n}$, the i-th row and the ij-th element of $\mathbf{M}$ are denoted by $\mathbf{m}_i$ and m_{ij}, respectively. The Frobenius norm of matrix $\mathbf{M}$ is denoted by $\|\mathbf{M}\|_F$. The trace of matrix $\mathbf{M}$ is denoted by $tr(\mathbf{M})$, and the transpose of matrix $\mathbf{M}$ is denoted by $\mathbf{M}^T$. $\mathbf{I}$ denotes an identity matrix, and $\mathbf{1}$ denotes a column vector with all the elements as one. The L2-norm and the L0-norm of vector $\mathbf{m}$ is denoted by $\|\mathbf{m}\|_2$ and $\|\mathbf{m}\|_0$, respectively.

2 New Clustering Formulations

To solve the three challenges mentioned in the introduction, we need to learn a new data graph $\mathbf{S}$ based on the given data graph $\mathbf{A}$ so that the graph $\mathbf{S}$ is more suitable for the final clustering task. In order to construct a clustering objective function based on this motivation, we start from the following theorem 1. Given an affinity matrix $\mathbf{S}$ which is nonnegative, then the Laplacian matrix $\mathbf{L}_s = \mathbf{D}_s - (\mathbf{S}^T + \mathbf{S})/2$ has the following important property.

Theorem 1 [Chung, 1997; Mohar *et al.*, 1991]: The multiplicity c of the eigenvalue zero of the Laplacian matrix $\mathbf{L}_s$ (nonnegative) is equal to the number of connected components in the graph with the similarity matrix $\mathbf{S}$.

Theorem 1 indicates that if $rank(\mathbf{L}_s) = n - c$, then the graph will be an ideal graph based on which we already partition the data points into c clusters, and there is no need to perform K-means or other discretization procedures.

Motivated by theorem 1, we aim to learn a similarity matrix $\mathbf{S} \in \mathbf{R}^{n\times n}$ (n is the number of samples) such that the corresponding Laplacian matrix $\mathbf{L}_s$ satisfied $rank(\mathbf{L}_s) = n - c$. Under this constraint, the learned $\mathbf{S}$ is block diagonal matrix, and we can directly partition the data points into c clusters based on $\mathbf{S}$ [Nie *et al.*, 2014]. To avoid the case that some rows of $\mathbf{S}$ are all zeros, we also constrain the $\mathbf{S}$ such that the sum of each row of $\mathbf{S}$ equals to one. Under all these constraints, we learn a similarity matrix $\mathbf{S}$ according to the following objective function.

$$\begin{aligned} &\min_{\mathbf{S}} \sqrt{\sum_{ij} \|\mathbf{a}_i - \mathbf{a}_j\|_2^2 s_{ij}} + \beta \|\mathbf{S}\|_F^2 \\ &s.t. \quad s_{ij} > 0, \quad \sum_j s_{ij} = 1, \quad rank(\mathbf{L}_s) = n - c \end{aligned} \tag{1}$$

The Lagrange function of Eq. (1) can be written as:

$$\sqrt{\sum_{ij} \|\mathbf{a}_i - \mathbf{a}_j\|_2^2 s_{ij}} + \beta \|\mathbf{S}\|_F^2 + \Gamma(\mathbf{\Lambda}, \mathbf{S}) \tag{2}$$

where $\mathbf{\Lambda}$ denotes the Lagrange multiplier, and $\Gamma(\mathbf{\Lambda}, \mathbf{S})$ denotes the formalized term derived from constraints. Taking the derivative of Eq. (2) w.r.t $\mathbf{S}$ and setting the derivative to zero, we get following equation:

$$v\frac{\partial \sum_{ij} \|\mathbf{a}_i - \mathbf{a}_j\|_2^2 s_{ij}}{\partial \mathbf{S}} + \frac{\partial \|\mathbf{S}\|_F^2}{\partial \mathbf{S}} + \frac{\partial \Gamma(\mathbf{\Lambda}, \mathbf{S})}{\partial \mathbf{S}} = 0 \tag{3}$$

where

$$v = \frac{1}{2\sqrt{\sum_{ij} \|\mathbf{a}_i - \mathbf{a}_j\|_2^2 s_{ij}}} \tag{4}$$

As v is related to the matrix $\mathbf{S}$, we cannot solve Eq. (3) directly. However, if v is set to be stationary, Eq. (3) can be considered accounting for following problem:

$$\begin{aligned} &\min_{\mathbf{S}} v \sum_{ij} \|\mathbf{a}_i - \mathbf{a}_j\|_2^2 s_{ij} + \beta \|\mathbf{S}\|_F^2 \\ &s.t. \quad s_{ij} > 0, \quad \sum_j s_{ij} = 1, \quad rank(\mathbf{L}_s) = n - c \end{aligned} \tag{5}$$

We can optimize Eq. (5) through an alternative method. Under the assumption that v is stationary, the Lagrange function of Eq. (2) can apply to Eq. (5). If $\mathbf{S}$ is solved by Eq. (5), the value of v can be updated correspondingly.

3 Optimization Algorithms

Let $\sigma_i(\mathbf{L}_s)$ denote the i-th smallest eigenvalue of $\mathbf{L}_s$. $\sigma_i(\mathbf{L}_s) \geq 0$ because $\mathbf{L}_s$ is positive semi-definite. Furthermore, if $\sum_{i=1}^{c} \sigma_i(\mathbf{L}_s) = 0$, the constraint condition $rank(\mathbf{L}_s) = n - c$ will be satisfied. According to Ky Fan's Theorem [Fan, 1949], we have:

$$v = \frac{1}{2\sqrt{\sum_{ij} \|\mathbf{a}_i - \mathbf{a}_j\|_2^2 s_{ij}}} \tag{6}$$

Therefore, the problem (5) can be equivalent to the following problem:

$$\begin{aligned} &\min_{\mathbf{S}} v \sum_{ij} \|\mathbf{a}_i - \mathbf{a}_j\|_2^2 s_{ij} + \beta \|\mathbf{S}\|_F^2 + 2\gamma tr(\mathbf{H}^T \mathbf{L}_s \mathbf{H}) \\ &s.t. \quad s_{ij} > 0, \quad \sum_j s_{ij} = 1, \quad \mathbf{H}^T\mathbf{H} = \mathbf{I} \end{aligned} \tag{7}$$

where γ is a very large number, and the optimal solution of the problem (7) can let equation $\sum_{i=1}^{c} \sigma_i(\mathbf{L}_s) = 0$ hold.

3.1 Fix S, Update v and H

When $\mathbf{S}$ is fixed, we can easily calculate the value of v by Eq. (4). So the first and second item of problem (7) can be seen as constants, then it transforms into:

$$\min_{\mathbf{H}\in\mathbf{R}^{n\times c}, \mathbf{H}^T\mathbf{H}=\mathbf{I}} tr(\mathbf{H}^T \mathbf{L}_s \mathbf{H}) \tag{8}$$

The optimal solution $\mathbf{H}$ is composed of the eigenvectors of $\mathbf{L}_s$ corresponding to the c smallest eigenvalues.

3.2 Fix v and H, Update S

When v and $\mathbf{H}$ are fixed, after a simple algebra, the Eq. (7) becomes:

$$\begin{aligned} &\min_{\mathbf{S}} \sum_{ij} (d_{ij} s_{ij} + \beta {s_{ij}}^2) + 2\gamma tr(\mathbf{H}^T \mathbf{L}_s \mathbf{H}) \\ &s.t. \quad s_{ij} > 0, \sum_j s_{ij} = 1 \end{aligned} \tag{9}$$

where $d_{ij} = v \|\mathbf{a}_i - \mathbf{a}_j\|_2^2$.

Note that there is an elementary but very important equation in spectral analysis

$$\sum_{i,j} \|\mathbf{h}_i - \mathbf{h}_j\|_2^2 s_{ij} = 2tr(\mathbf{H}^T \mathbf{L}_s \mathbf{H}) \tag{10}$$

Denote $e_{ij} = \|\mathbf{h}_i - \mathbf{h}_j\|_2^2$, since the Eq. (9) is independent between different i, we can solve following problem individually for each i:

$$\begin{aligned} &\min_{\mathbf{S}} \sum_{ij} \left(d_{ij} s_{ij} + \beta {s_{ij}}^2 + \gamma e_{ij} s_{ij}\right) \\ &s.t. \;\; s_{ij} > 0, \;\; \sum_j s_{ij} = 1 \end{aligned} \tag{11}$$

Let $\mathbf{f}_i \in \mathbf{R}^{n\times 1}$ denote the vector with the j-th element as $f_{ij} = d_{ij} + \lambda e_{ij}$, then Eq. (11) is rewritten as follow:

$$\min_{\mathbf{s}_i} \left\| \mathbf{s}_i + \frac{1}{2\beta}\mathbf{f}_i \right\|_2^2 \;\; s.t. \;\; s_{ij} > 0, \sum_j s_{ij} = 1 \tag{12}$$

The problem (12) can be solved by an efficient iterative algorithm [Huang *et al.*, 2015] and the intermediate variable β can be determined using the number of adaptive neighbours [Nie *et al.*, 2017]. By iteratively solving problem (5), we can get the final $\mathbf{S}$ in the objective function Eq. (1). The pseudo code is summarized in Algorithm 1. In this algorithm, we update the m nearest similarities for each sample in $\mathbf{S}$ and thus the complexity of updating $\mathbf{S}$ and $\mathbf{H}$ (only requires computing the top c eigenvectors of a very sparse matrix) is reduced significantly. In our further work, we will make this technique reliable on very large scale datasets.

Algorithm 1:

Input: affinity matrix $\mathbf{A} \in \mathbf{R}^{n\times n}$, cluster number c, parameter γ.

Initialize $v = 1$, then each row $\mathbf{s}_i$ of $\mathbf{S}$ can be initialized by solving the following problem:

$$\min_{s_{ij}>0, \sum_j s_{ij}=1} \sum_{j=1}^{n} \left(\|\mathbf{a}_i - \mathbf{a}_j\|_2^2 s_{ij} + \beta {s_{ij}}^2 \right)$$

repeat

1. Update v by using Eq. (4).
2. Update $\mathbf{H}$ by solving the Eq. (8).
3. Update each row of $\mathbf{S}$ by solving the Eq. (12).

until converge

Output: similarity matrix $\mathbf{S} \in \mathbf{R}^{n\times n}$ with exact c connected components.

4 Convergence Analysis

The proposed algorithm 1 can find a local optimal solution. Before proving the convergence, we first introduce the following lemma [Marsden *et al.*, 1993].

Lemma 1: For any positive real number w and q, the following inequality holds:

$$2\sqrt{w}\sqrt{q} \le w + q \tag{13}$$

Theorem 2: In Algorithm 1, updated $\mathbf{S}$ will decrease the objective value of problem (1) until converge.

Proof: Suppose the updated $\mathbf{S}$ is $\widehat{\mathbf{S}}$ in each iteration, so $\widehat{\mathbf{S}}$ makes the objective of Eq. (5) have the smaller value than $\mathbf{S}$. Combining Eq. (4), we have

$$\begin{aligned} &\frac{\sum_{ij} \|\mathbf{a}_i - \mathbf{a}_j\|_2^2 \widehat{s_{ij}}}{2\sqrt{\sum_{ij} \|\mathbf{a}_i - \mathbf{a}_j\|_2^2 s_{ij}}} + \beta \left\|\widehat{\mathbf{S}}\right\|_F^2 \\ &\le \frac{\sum_{ij} \|\mathbf{a}_i - \mathbf{a}_j\|_2^2 s_{ij}}{2\sqrt{\sum_{ij} \|\mathbf{a}_i - \mathbf{a}_j\|_2^2 s_{ij}}} + \beta \|\mathbf{S}\|_F^2 \end{aligned} \tag{14}$$

According to Lemma 1, we have

$$\begin{aligned} &2\sqrt{\sum_{ij} \|\mathbf{a}_i - \mathbf{a}_j\|_2^2 \widehat{s_{ij}}} \sqrt{\sum_{ij} \|\mathbf{a}_i - \mathbf{a}_j\|_2^2 s_{ij}} \\ &\le \sum_{ij} \|\mathbf{a}_i - \mathbf{a}_j\|_2^2 \widehat{s_{ij}} + \sum_{ij} \|\mathbf{a}_i - \mathbf{a}_j\|_2^2 s_{ij} \end{aligned} \tag{15}$$

after a simple algebra, the Eq. (15) becomes:

$$\begin{aligned} &\sqrt{\sum_{ij} \|\mathbf{a}_i - \mathbf{a}_j\|_2^2 \widehat{s_{ij}}} - \frac{\sum_{ij} \|\mathbf{a}_i - \mathbf{a}_j\|_2^2 \widehat{s_{ij}}}{2\sqrt{\sum_{ij} \|\mathbf{a}_i - \mathbf{a}_j\|_2^2 s_{ij}}} \\ &\le \sqrt{\sum_{ij} \|\mathbf{a}_i - \mathbf{a}_j\|_2^2 s_{ij}} - \frac{\sum_{ij} \|\mathbf{a}_i - \mathbf{a}_j\|_2^2 s_{ij}}{2\sqrt{\sum_{ij} \|\mathbf{a}_i - \mathbf{a}_j\|_2^2 s_{ij}}} \end{aligned} \tag{16}$$

According to Eq. (14) and Eq. (16), we have:

$$\begin{aligned} &\sqrt{\sum_{ij} \|\mathbf{a}_i - \mathbf{a}_j\|_2^2 \widehat{s_{ij}}} + \beta \left\|\widehat{\mathbf{S}}\right\|_F^2 \\ &\le \sqrt{\sum_{ij} \|\mathbf{a}_i - \mathbf{a}_j\|_2^2 s_{ij}} + \beta \|\mathbf{S}\|_F^2 \end{aligned} \tag{17}$$

■

5 Learning An Initial Graph

In the algorithm 1, before learning the normalized and block-diagonal similarity matrix $\mathbf{S} \in \mathbf{R}^{n\times n}$. an initial graph affinity matrix $\mathbf{A}$ is required to be given. In this section, similar with [Nie *et al.*, 2016], we give an approach to initialize graph $\mathbf{A}$. We want to learn a normalized and nonnegative similarity matrix $\mathbf{S}$, so it is desirable for the initial graph $\mathbf{A}$ to have the same constraint.

Given the data points $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$, we want to learn the affinity values of A such that smaller distance $\|\mathbf{x}_i - \mathbf{x}_j\|_2^2$ between data points $\mathbf{x}_i$ and $\mathbf{x}_j$ corresponds to a larger affinity value a_{ij}. Then we solve the following problem:

$$\min_{a_{ij}\ge 0,\ \mathbf{a}_i^T \mathbf{1}=1,\ a_{ii}=0} \sum_{j=1}^{n} z_{ij} a_{ij} + \mu \sum_{j=1}^{n} a_{ij}^2 \tag{18}$$

where $z_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2$, furthermore, in many cases, we want to get a sparse affinity matrix A for higher performance. Therefore, we learn the affinities with the maximal u such that the optimal solution $\mathbf{a}_i$ to the question (18) has exactly m nonzero values; i.e., the L0-norm of $\mathbf{a}_i$ is constrained to be m. To this end, we solve the following problem:

$$\max_{\mu, \left\|\widehat{\mathbf{a}}_i\right\|_0 = m} \mu \tag{19}$$

where $\widehat{\mathbf{a}}_i$ is the optimal solution to the Eq. (18).

After a simple algebra, Eq. (18) becomes:

$$\min_{a_{ij}\geq 0, \mathbf{a}_i^T\mathbf{1}=1, a_{ii}=0} \frac{1}{2}\left\|\mathbf{a}_i + \frac{\mathbf{z}_i}{2\mu}\right\|_2^2 \tag{20}$$

The Lagrangian function of Eq. (20) is:

$$\xi(\mathbf{a}_i, \delta, \eta_i) = \frac{1}{2}\left\|\mathbf{a}_i + \frac{\mathbf{z}_i}{2\mu}\right\|_2^2 - \delta(\mathbf{a}_i^T\mathbf{1} - 1) - \eta_i^T\mathbf{a}_i \tag{21}$$

where δ and η_i are the Lagrange multipliers.

The optimal solution $\widehat{\mathbf{a}}$ should satisfy that the derivative of Eq. (21) w.r.t. $\mathbf{a}_i$ is equal to zero, so we have

$$\widehat{a}_{ij} + \frac{z_{ij}}{2\mu} - \delta - \eta_{ij} = 0 \tag{22}$$

Noting that $a_{ij}\eta_{ij} = 0$ and according to the KKT condition, we have

$$\widehat{a}_{ij} = \left(-\frac{z_{ij}}{2\mu} + \delta\right)_+ \tag{23}$$

According to the constraint $\left\|\widehat{\mathbf{a}}_i\right\|_0 = m$ in Eq. (19), we know $\widehat{a}_{im} > 0$ and $\widehat{a}_{i,m+1} > 0$. Therefore, we have

$$\begin{cases} \frac{z_{im}}{2\mu} + \delta > 0 \\ \frac{z_{i,m+1}}{2\mu} + \delta \leq 0 \end{cases} \tag{24}$$

According to Eq. (23) and the constraint in Eq. (18), we have

$$\delta = \frac{1}{m} + \frac{1}{2m\mu}\sum_{j=1}^{m} z_{ij} \tag{25}$$

According to Eq. (24) and Eq. (25), we can get the following inequality:

$$\frac{m}{2}z_{im} - \frac{1}{2}\sum_{j=1}^{m} z_{ij} < \mu < \frac{m}{2}z_{i,m+1} - \frac{1}{2}\sum_{j=1}^{m} z_{ij} \tag{26}$$

Therefore, to obtain the optimal solution $\widehat{\mathbf{a}}_i$ to the Eq. (18) that has exactly m nonzero values, the maximal is

$$\mu = \frac{m}{2}z_{i,m+1} - \frac{1}{2}\sum_{j=1}^{m} z_{ij} \tag{27}$$

According to Eqs. (23), (25) and (27), we have the optimal affinities $\widehat{a}_{ij}$ as follows:

$$\widehat{a}_{ij} = \begin{cases} \frac{z_{i,m+1} - z_{ij}}{m z_{i,m+1} - \sum_{k=1}^{m} z_{ik}} & j \leq m \\ 0 & j > m \end{cases} \tag{28}$$

The affinities $\widehat{a}_{ij}$ computed by Eq. (28) have the following advantages [Nie *et al.*, 2016]:

(1). Eq. (28) only involves the basic operations (i.e. addition, subtraction, multiplication and division). Some methods often can be used to compute the affinities such as LLE [Roweis and Saul, 2000] and sparse coding [Wright *et al.*, 2009], but the computations of Gaussian functions and other more operations that make these methods less efficient than our method.

(2). The learned Initial matrix $\widehat{\mathbf{A}}$ is naturally sparse, which is computationally efficient for graph-based learning tasks such as clustering.

(3). The property that the affinities are distance consistent is guaranteed from the motivation of this method. If the distance between samples $\mathbf{x}_i$ and$\mathbf{x}_j$ is smaller than the distance between $\mathbf{x}_i$ and $\mathbf{x}_k$, then the affinity $\widehat{a}_{ij}$ computed by Eq. (28) is larger than the affinity $\widehat{a}_{ik}$. However, this property is not guaranteed in LLE and sparse coding.

(4). Computing the affinities by Eq. (28) only involves one parameter i.e., the number of neighbors m. This parameter is easy to tune. In general, $m < 10$ can produce good results. This property is important because the tuning of parameters remains a open and difficult problem in clustering task. In graph-based clustering task, there are few labeled samples and thus traditional parameter tuning techniques such as cross validation will not be used.

6 Experiments

In this section, we will validate our proposed method on four benchmark datasets (Handwritten numerals, MSRC-v1, COIL20 and UMIST), and compare it with other clustering methods: spectral clustering (SC) [Ng *et al.*, 2002], K-means clustering [MacQueen and others, 1967], Constrained Laplacian Rank (CLR) [Nie *et al.*, 2016]. In our model, we determine the value of γ in a heuristic way to accelerate the procedure. At first, we set γ with a small positive value, then in each iteration, decrease it ($\gamma = \gamma/2$) if the number of zero eigenvalues in $\mathbf{L}_s$ is larger than class number c or increase it ($\gamma = 2\gamma$) if smaller than c, otherwise the iteration stopped.

6.1 Datasets Descriptions

COIL20 databast [Nene *et al.*, 1996] includes 1440 color images of 20 objects (72 images per object). The object has a wide variety of complex geometric and reflectance characteristics. This database, called Columbia Object Image Library (COIL-20), was used in a real-time 20 object recognition system. Each object was placed in a stable configuration at approximately the center of the turntable. The turntable was then rotated through 360 degrees and 72 images were taken per object; one at every 5 degrees of rotation.

UMIST dataset [Graham and Allinson, 1998] consists of 564 images of 20 individuals (mixed race/gender/appearance). Each individual is shown in a range of poses from profile to frontal views and images are numbered consecutively as they were taken. The files are all in PGM format, approximately 220×220 pixels with 256-bit grey-scale.

Handwritten numerals (HW) dataset [Asuncion and Newman, 2007] is composed of 2,000 data points for 0 to 9 ten digit classes and each class has 200 data points. Five published features can be used for clustering: View1: 76 Fourier coefficients of the character shapes (FOU), View2: 216 profile correlations (FAC), View3: 240 pixel averages in 2×3 windows (PIX), View4: 47 Zernike moment (ZER) and View5: 6 morphological (MOR) features.

MSRC-v1 dataset [Winn and Jojic, 2005] contains 240 images and can be divided into 8 classes. Following [Lee and

		View 1	View 2	View 3	View 4	View 5
ACC	SC	53.49	56.88	95.19	55.85	15.31
	K-means	60.89	60.06	72.72	53.00	39.11
	CLR	61.02	66.56	89.79	58.79	41.11
	Proposed	**65.60**	**71.37**	**96.91**	**64.33**	**44.22**
NMI	SC	62.09	58.88	90.93	58.54	10.83
	K-means	62.51	60.06	71.79	49.91	46.78
	CLR	66.83	70.16	88.56	60.39	45.16
	Proposed	**68.69**	**74.11**	**93.09**	**65.01**	**47.28**

Table 1: The ACC(%) and NMI(%) of four methods under different views on HW dataset.

Grauman, 2009], we select 7 classes composed of tree, building, airplane, cow, face, car, bicycle and each class has 30 images. To distinguish all of scenes, we extract four visual features from each image: View1: 24 Color Moment, View2: 576 Histogram of Oriented Gradient, View3: 256 Local Binary Pattern and View4: 254 Centrist features.

6.2 Evaluation Criteria

In experiments, we evaluate the clustering performance with two standard clustering evaluation metrics, i.e. Accuracy (ACC) [Cai *et al.*, 2005] and Normalized Mutual Information (NMI) [Estévez *et al.*, 2009].

Given an input data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots \mathbf{x}_n] \in \mathbf{R}^{d\times n}$ composed of n feature vectors of d dimensions as its columns, $\mathbf{r} = [r_1, r_2, \cdots r_n] \in \mathbf{R}^n$ and $\tilde{\mathbf{r}} = [\tilde{r}_1, \tilde{r}_2, \cdots \tilde{r}_n] \in \mathbf{R}^n$ denote the truth class label vector and predict class label vector, respectively. r_i, and $\tilde{r}_i$ are the truth label and predict label of $\mathbf{x}_i$.

Accuracy (ACC): ACC measures one-to-one relationship between predict class and the true classes:

$$ACC = \frac{\sum_{i=1}^{n} \delta(\tilde{r}_i, map(r_i))}{n} \tag{29}$$

where $map(r_i)$ is a permutation mapping function that maps each cluster label r_i to the equivalent label from the dataset. $\delta(x, y)$ is the delta function defined as:

$$\delta(x, y) = \begin{cases} 1, & x = y \\ 0, & otherwise \end{cases} \tag{30}$$

Normalized Mutual Information (NMI): NMI measures the similarity between the clustering results and the true classes:

$$NMI(\mathbf{r}, \tilde{\mathbf{r}}) = \frac{MI(\mathbf{r}, \tilde{\mathbf{r}})}{\sqrt{H(\mathbf{r})H(\tilde{\mathbf{r}})}} \tag{31}$$

where Mutual Information $MI(\mathbf{r}, \tilde{\mathbf{r}})$ and entropy $H(\mathbf{r})$ are respectively calculated as follows:

$$MI(\mathbf{r}, \tilde{\mathbf{r}}) = \sum_{r_i \in \mathbf{r}, \tilde{r}_j \in \tilde{\mathbf{r}}} p(r_i, \tilde{r}_j) \cdot \log_2 \frac{p(r_i, \tilde{r}_j)}{p(r_i)p(\tilde{r}_j)} \tag{32}$$

$$H(\mathbf{r}) = -\sum_{r_i \in \mathbf{r}} p(r_i) \cdot \log_2 p(r_i) \tag{33}$$

where $p(r_i)$ is the probabilities that the arbitrary image belongs to the cluster r_i, and $p(r_i, \tilde{r}_j)$ is the joint probability that the image belongs to the both the clusters r_i and cluster $\tilde{r}_j$.

6.3 Experimental Results Analysis

For each dataset, we repeat experiments 10 times because all the methods are spectral clustering based methods. Spectral clustering has k-means procedure, however, the performance of it has large relationship with the choice of initial centroids. Finally, we report the average performance including ACC and NMI. Table 1 and Table 2 show the results on HW dataset and MSRC-v1 dataset, respectively. Table 3 shows the results on COIL20 dataset and UMIST dataset. Figure 1 shows the converge curve of our method on four datasets. Comparing the aforementioned experimental results, we have several interesting observations as follows:

(1) From Table 1 and Table 2 we conclude that our proposed methods outperform the competing methods on each view, which means our model is robust to different visual features on clustering task. The results of Table 3 also shows our model obtains the best performance on COIL20 dataset and UMIST dataset. Moreover, our model is very robust to the parameter γ, it nearly can be seen parameter-free approach. By contrast, the parameter-free method CLR is robust to some extent, but the performance is not satisfactory.

(2) In order to show the convergence effect on the same graph, we choose the log value of the objective function as the ordinate. Figure 1 shows the log value of function converges within only three steps. This is consistent with our analysis in Section 3. It illustrates that our algorithm has a good practical application for its good convergence.

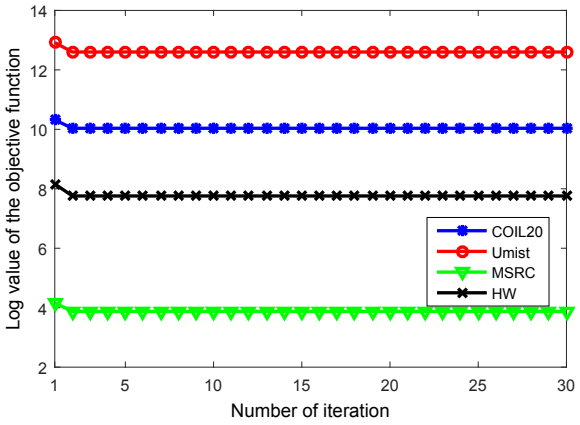

Figure 1: Convergence curve of our model on four datasets.

		View 1	View 2	View 3	View 4
ACC	SC	36.51	65.82	50.74	63.55
	K-means	35.45	60.90	49.74	53.28
	CLR	32.28	50.00	47.88	54.55
	Proposed	**37.14**	**73.33**	**58.12**	**71.96**
NMI	SC	28.75	60.80	44.80	53.77
	K-means	25.06	53.45	41.99	45.57
	CLR	24.18	47.15	45.19	48.39
	Proposed	**29.26**	**64.60**	**46.00**	**60.13**

Table 2: The ACC(%) and NMI(%) of four methods under different views on MSRC-v1 dataset.

		COIL20	UMIST
ACC	SC	75.83	62.40
	K-means	53.48	42.46
	CLR	80.08	50.47
	Proposed	**81.73**	**65.52**
NMI	SC	86.03	77.99
	K-means	69.78	62.20
	CLR	87.76	72.89
	Proposed	**89.45**	**79.64**

Table 3: The ACC(%) and NMI(%) of four methods under different views on COIL20 dataset and UMIST dataset.

7 Conclusions

In this paper, a novel graph-based clustering model is proposed to learn a new graph with exactly c connected components (where c is the number of clusters). Instead of fixing the original data graph associated to the affinity matrix, the proposed model tries to learn a new block diagonal data similarity matrix such that the clustering results can be immediately obtained without requiring any post-processing to extract the clustering indicators. An optimization algorithm is also given to solve our model. Empirical results on some real benchmark datasets showed our method is significantly better than several well-established clustering approaches.

Acknowledgements

This work is supported by National Natural Science Foundation of China under Grant 61773302 and the 111 Project of China (B08038).

References

[Asuncion and Newman, 2007] Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

[Ben-Hur *et al.*, 2001] Asa Ben-Hur, David Horn, Hava T Siegelmann, and Vladimir Vapnik. Support vector clustering. *Journal of machine learning research*, 2(Dec):125–137, 2001.

[Cai *et al.*, 2005] Deng Cai, Xiaofei He, and Jiawei Han. Document clustering using locality preserving indexing. *IEEE Transactions on Knowledge and Data Engineering*, 17(12):1624–1637, 2005.

[Cai *et al.*, 2013] Xiao Cai, Feiping Nie, and Heng Huang. Multi-view k-means clustering on big data. In *IJCAI*, pages 2598–2604, 2013.

[Chung, 1997] Fan RK Chung. *Spectral graph theory*. Number 92. American Mathematical Soc., 1997.

[Ding *et al.*, 2005] Chris Ding, Xiaofeng He, and Horst D Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, pages 606–610. SIAM, 2005.

[Elhamifar and Vidal, 2013] Ehsan Elhamifar and Rene Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2765–2781, 2013.

[Estévez *et al.*, 2009] Pablo A Estévez, Michel Tesmer, Claudio A Perez, and Jacek M Zurada. Normalized mutual information feature selection. *IEEE Transactions on Neural Networks*, 20(2):189–201, 2009.

[Fan, 1949] Ky Fan. On a theorem of weyl concerning eigenvalues of linear transformations i. *Proceedings of the National Academy of Sciences*, 35(11):652–655, 1949.

[Gao *et al.*, 2018] Quanxue Gao, Lan Ma, Yang Liu, Xinbo Gao, and Feiping Nie. Angle 2dpca: A new formulation for 2dpca. *IEEE Transactions on Cybernetics*, 48(5):1672–1678, 2018.

[Graham and Allinson, 1998] Daniel B Graham and Nigel M Allinson. Characterising virtual eigensignatures for general purpose face recognition. In *Face Recognition*, pages 446–456. Springer, 1998.

[Hagen and Kahng, 1992] Lars Hagen and Andrew B Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE transactions on computer-aided design of integrated circuits and systems*, 11(9):1074–1085, 1992.

[Huang *et al.*, 2015] Jin Huang, Feiping Nie, and Heng Huang. A new simplex sparse learning model to measure data similarity for clustering. In *IJCAI*, pages 3569–3575, 2015.

[Johnson, 1967] Stephen C Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[Lee and Grauman, 2009] Yong Jae Lee and Kristen Grauman. Foreground focus: Unsupervised learning from partially matching images. *International Journal of Computer Vision*, 85(2):143–166, 2009.

[Liu *et al.*, 2013] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2013.

[Liu *et al.*, 2018] Yang Liu, Kaiwen Wen, Quanxue Gao, Xinbo Gao, and Feiping Nie. Svm based multi-label learning with missing labels for image annotation. *Pattern Recognition*, 78:307–317, 2018.

[MacQueen and others, 1967] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA., 1967.

[Marsden *et al.*, 1993] SS Antman JE Marsden, L Sirovich S Wiggins, L Glass, RV Kohn, and SS Sastry. *Interdisciplinary Applied Mathematics*. Springer, 1993.

[Mohar *et al.*, 1991] Bojan Mohar, Y Alavi, G Chartrand, and OR Oellermann. The laplacian spectrum of graphs. *Graph theory, combinatorics, and applications*, 2(871-898):12, 1991.

[Nene *et al.*, 1996] Sameer A Nene, Shree K Nayar, and Hiroshi Murase. Columbia object image library (coil-20)(technical report). *Columbia University*, 1996.

[Ng *et al.*, 2002] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

[Nie *et al.*, 2011] Feiping Nie, Zinan Zeng, Ivor W Tsang, Dong Xu, and Changshui Zhang. Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering. *IEEE Transactions on Neural Networks*, 22(11):1796–1808, 2011.

[Nie *et al.*, 2014] Feiping Nie, Xiaoqian Wang, and Heng Huang. Clustering and projected clustering with adaptive neighbors. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 977–986. ACM, 2014.

[Nie *et al.*, 2016] Feiping Nie, Xiaoqian Wang, Michael I Jordan, and Heng Huang. The constrained laplacian rank algorithm for graph-based clustering. In *AAAI*, pages 1969–1976, 2016.

[Nie *et al.*, 2017] Feiping Nie, Guohao Cai, and Xuelong Li. Multi-view clustering and semi-supervised classification with adaptive neighbours. In *AAAI*, pages 2408–2414, 2017.

[Roweis and Saul, 2000] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[Shi and Malik, 2000] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

[Wang *et al.*, 2013] Hua Wang, Feiping Nie, and Heng Huang. Multi-view clustering and feature learning via structured sparsity. In *International Conference on Machine Learning*, pages 352–360, 2013.

[Wang *et al.*, 2017] Qianqian Wang, Lan Ma, Quanxue Gao, Yunsong Li, Yunfang Huang, and Yang Liu. Adaptive maximum margin analysis for image recognition. *Pattern Recognition*, 61:339–347, 2017.

[Winn and Jojic, 2005] John Winn and Nebojsa Jojic. Locus: Learning object classes with unsupervised segmentation. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 756–763. IEEE, 2005.

[Wright *et al.*, 2009] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence*, 31(2):210–227, 2009.

[Xu *et al.*, 2005] Linli Xu, James Neufeld, Bryce Larson, and Dale Schuurmans. Maximum margin clustering. In *Advances in neural information processing systems*, pages 1537–1544, 2005.

[Zass and Shashua, 2007] Ron Zass and Amnon Shashua. Doubly stochastic normalization for spectral clustering. In *Advances in Neural Information Processing Systems*, pages 1569–1576, 2007.