The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

New $\ell_{2,1}$-Norm Relaxation of Multi-Way Graph Cut for Clustering

Xu Yang,[1] Cheng Deng,[1] Xianglong Liu,[2]* Feiping Nie[3]
[1]School of Electronic Engineering, Xidian University, Xian 710071, China
[2]Beihang University, Beijing 100191, China
[3]Northwestern Polytechnical University, Xian 710072, China
{xuyang.xd, chdeng.xd, mldeeper}@gmail.com , xlliu@nlsde.buaa.edu.cn

Abstract

The clustering methods have absorbed even-increasing attention in machine learning and computer vision communities in recent years. Exploring manifold information in multi-way graph cut clustering, such as ratio cut clustering, has shown its promising performance. However, traditional multi-way ratio cut clustering method is NP-hard and thus the spectral solution may deviate from the optimal one. In this paper, we propose a new relaxed multi-way graph cut clustering method, where $\ell_{2,1}$-norm distance instead of squared distance is utilized to preserve the solution having much more clearer cluster structures. Furthermore, the resulting solution is constrained with normalization to obtain more sparse representation, which can encourage the solution to contain more discrete values with many zeros. For the objective function, it is very difficult to optimize due to minimizing the ratio of two non-smooth items. To address this problem, we transform the objective function into a quadratic problem on the Stiefel manifold (QPSM), and introduce a novel yet efficient iterative algorithm to solve it. Experimental results on several benchmark datasets show that our method significantly outperforms several state-of-the-art clustering approaches.

Introduction

As an important task in machine learning and computer vision communities, clustering has been widely used in image segmentation (Shi and Malik 2000), image categorization (Grauman and Darrell 2006), and digital media analysis (An et al. 2012). The goal of clustering is to find a partition in order to keep similar data vectors in the same cluster while those dissimilar ones in different clusters. In recent years, many clustering methods have been proposed, such as K-means clustering, spectral clustering, support vector clustering (Ben-Hur et al. 2001), and non-negative matrix factorization clustering (Xu, Liu, and Gong 2003). Among these methods, exploring manifold information in spectral clustering can greatly promote the clustering performance, whose basic idea is to transform the clustering problem into the optimal partition problem of a graph via spectral theory. The graph-based learning aims at capturing the intrinsic manifold structure underlying the images (Deng et al. 2013; 2014) and the relationship of features (Liu, Tsang, and Müller 2017) to obtain better results. The clustering methods based on graph cut exploit a weighted undirected graph to represent the similarities of the data, and try to find the best cuts of the graph to produce the ideal clustering results. However, finding such optimal graph cut is a NP-hard problem. To address this problem, the graph cut clustering problem is generally relaxed into the eigenvector space of continuous Laplace matrix, and different segmentations corresponds to different eigenvector spaces.

The graph cut clustering mainly contains two-way partitioning and multi-way partitioning. Two-way partitioning divides a graph into exactly two subgraphs, while multi-way partitioning divides a graph into multiple vertex disjoint subgraphs. Any two-way method can be applied to generate multiple partitions, but the quality of the solution may be unsatisfied due to the inherent limitation of the result. Hence, many multi-way clustering methods have been successively proposed by using the multi-dimensional spectral embedding. The ratio cut clustering (Chan, Schlag, and Zien 1994) and the normalized cut clustering (Ng, Jordan, and Weiss 2002) are the two most common multi-way spectral clustering methods, in which Laplacian embedding and the generalized eigenvectors of Laplace matrix are respectively used to produce the optimal solution. Practically, these solutions can not be directly regarded as clustering results, and a further clustering algorithm such as K-means has to be adopted to obtain the final clustering results. However, this traditional relaxation strategy is more likely to cause the results deviating from the optimal solution. More recently, tight relaxation of balanced graph clustering method was proposed (Nie et al. 2016a), where the solution is a vector to partition the data affinity graph into two disjoint sets. It requires to recursively run the clustering method to obtain the desired number of partitions, which can only get low-quality approximated multi-way clustering results and also leads to the high computational cost.

In this paper, we propose a new relaxation for the multi-way graph cut clustering method to obtain much clearer cluster structures which solve the multi-way clustering task without approximation. Different from traditional ratio cut graph clustering method, we constrain the normalized solution with the $\ell_{2,1}$-norm to enforce each element to have equal chance to be zero. Thus, the solution of our new

*Corresponding author.

relaxed graph clustering method can obtain discrete values with many zeros, which approximates the ideal solution. However, our proposed clustering method is difficult to optimize because of including two non-smooth proportional items. To deal with it, two different treatments are composted in our optimization, and we then transform the objective function into a quadratic problem on the Stiefel manifold (QPSM) (Nie, Zhang, and Li 2017), which takes much less time for the convergence. Empirical experiments demonstrate our method consistently achieves better clustering results than several state-of-the-art clustering methods in both synthetic data and nine real datasets.

Related Work

Suppose we have *n* data points $X = [x_1, x_2, ..., x_n] \in \mathbf{R}^{n\times d}$ and weight matric W. Our main propose is to divide these data points into different disjoint clusters and the data points within each cluster are adjacent.

Ratio Cut in Two-Way Clustering

The ratio cut in two-way clustering (Cheng and Wei 1991; Hagen and Kahng 1992) is defined as:

$$Rcut = \frac{cut(A,B)}{|A|} + \frac{cut(A,B)}{|B|}, \tag{1}$$

where $|A|$ represents the number of data points in A, similarly in B. $cut(A,B) = \sum_{i\in A, j\in B} W_{ij}$.

Denote an indictor vector $y \in \mathbf{R}^{n\times 1}$ as follows:

$$y = [\underbrace{1, ..., 1}_{n_1}, \underbrace{r, ..., r}_{n_2}]^\top,$$

where $n_1 + n_2 = n$, $y_i = 1$ if data point i is in A, otherwise $y_i = r$.

As suggested in (Shi and Malik 2000), we can know that when $r = -\frac{n_1}{n_2}$, the ratio cut can be written as:

$$Rcut = \frac{\frac{1}{2}\sum_{i,j} W_{ij}|y_i - y_j|^2}{\sum_i |y_i|^2} = \frac{y^\top L y}{y^\top y}, \tag{2}$$

where $L = D - W$ is the Laplacian matrix, and D is the diagonal matrix with $D_{ii} = \sum_j W_{ij}$. When we need to minimize the normalized cut to obtain an ideal partition, Eq. (2) can be devised as:

$$\min_{y=[1,...,1,-\frac{n_1}{n_2},...,-\frac{n_1}{n_2}]^\top} \frac{\frac{1}{2}\sum_{i,j} W_{ij}|y_i - y_j|^2}{\sum_i |y_i|^2}. \tag{3}$$

This is a NP-hard problem when we want to obtain the best cut value, so we can write constraints as $y^\top \mathbf{1} = 0$, where $\mathbf{1}$ is used to denote a column vector with all elements being one. Then, Eq. (3) can be formulated as:

$$\min_{y^\top \mathbf{1}=0} \frac{\frac{1}{2}\sum_{i,j} W_{ij}|y_i - y_j|^2}{\sum_i |y_i|^2}. \tag{4}$$

The optimal solution y is formed by the eigenvector of L corresponding to the second smallest eigenvector. In order to avoid the solution y deviating from the constraint $y^\top \mathbf{1} = 0$, we can take 0 or the median value as the splitting point to obtain the best normalized cut value (Shi and Malik 2000).

Ratio Cut in Multi-Way Clustering

A two-way partitioning divides a graph into exactly two subgraphs, whereas multi-way partitioning divides the graph into multiple vertex disjoint subgraphs. We convert the two-way ratio-cut indictor vector to a multi-way ratio-cut indictor metric $F \in \mathbf{R}^{n\times c}$, where f_k is the k-th column of F and c is the number of clusters. The multi-way graph ratio cut clustering based on the multi-dimensional spectral embedding (Chan, Schlag, and Zien 1994) is defined as:

$$\min_{F^\top F = I} \sum_{i,j}^{n} W_{ij}\|f_i - f_j\|_2^2. \tag{5}$$

It is obvious that minimizing Eq. (5) is NP-hard, and traditional method solve the ratio clustering with relaxation:

$$\min_{F^\top F = I} Tr(F^\top L F). \tag{6}$$

The optimal solution F is formed by the c eigenvectors of L corresponding to the c smallest eigenvalues. The solution F of the relaxed problems can not be directly used as clustering results, and a further clustering algorithm such as K-means has to be utilized on F to obtain the final clustering results.

New $\ell_{2,1}$-Norm Relaxation of Multi-Way Graph Cut for Clustering

The obtained solution of traditional spectral relaxation in multi-way graph cut may deviate from the ideal solution. In this section, we propose a new relaxation for ratio cut in multi-way clustering to make the solution are approximate to optimal ones. We constrain the normalized solution with the $\ell_{2,1}$-norm to obtain much clear cluster structures. Moreover, a new iterative optimization algorithm is adopted to address our proposed method.

Under the constraints $y^\top y = 1$, we can prove that Eq. (4) is equivalent to:

$$\min_{y^\top y=1, y^\top \mathbf{1}=0} \frac{\frac{1}{2}\sum_{i,j} W_{ij}|y_i - y_j|^2}{\sum_{i,j} |y_i - y_j|^2}, \tag{7}$$

where $y_i = \sqrt{\frac{n_2}{n_1^2 + n_1 n_2}}$ if data point i is in A, otherwise $y_i = -\sqrt{\frac{n_1}{n_2^2 + n_1 n_2}}$. It shows that clustering keeps similar vectors in the same cluster while those dissimilar ones in different clusters. Thus, it naturally extends to multi-way situations:

$$\min_{F^\top F = I, F^\top \mathbf{1}=0} \frac{\sum_{i,j} W_{ij}\|f_i - f_j\|_2^2}{\sum_{i,j} \|f_i - f_j\|_2^2}, \tag{8}$$

where $f_i \in \mathbf{R}^{1\times c}$ is the indictor vector of the data x_i. In the actual process, the feature vectors of the Laplacian matrix is adopted to complete the segmentation. Hence, Eq. (8) can be converted to:

$$\min_{F^\top F = I, F^\top \mathbf{1}=0} \frac{Tr(F^\top L F)}{Tr(F^\top H F)}, \tag{9}$$

where $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. Under the constraint $F^\top \mathbf{1} = 0$, it can be easily proved that:

$$\begin{aligned} Tr(F^\top HF) &= Tr(F^\top (I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)F) \\ &= Tr(F^\top F) - \frac{1}{n}Tr(F^\top(\mathbf{1}\mathbf{1}^\top)F) = c. \end{aligned} \tag{10}$$

Therefore, when the loss function is a square loss, Eq. (9) is equivalent to the traditional multi-way ratio cut clustering algorithm:

$$\min_{F^\top F=I, F^\top \mathbf{1}=0} Tr(F^\top LF). \tag{11}$$

Denote $\|F\|_{2,1} = \sum_i \|f_i\|_2$ as the $\ell_{2,1}$-norm of matrix F. The best solution F for clustering is that $f_i = f_j$ if x_i and x_j belong to the same cluster. That is to say, $\|f_i - f_j\|_2 = 0$ for many pairs of (i, j) in same cluster. In addition, the square loss function is very sensitive to outliers. Hence, the robust $\ell_{2,1}$-norm based loss function is adopted in our method to obtain the ideal solution. The objective of our method can be formulated as:

$$\min_{F^\top F=I, F^\top \mathbf{1}=0} \frac{\sum_{i,j} W_{ij}\|f_i - f_j\|_2}{\sum_{i,j}\|f_i - f_j\|_2}. \tag{12}$$

It is widely known that minimizing $\ell_{2,1}$-norm usually generates sparse solutions (Nie et al. 2010; 2011). That is to say, the solution F will take on discrete values and have more zero elements. Moreover, Eq. (12) can keep the distance of indicator vectors similar if data belongs to the same cluster, even make them equal. The distance of indicator vectors as separated as possible if data belongs to the different clusters. Thus, the proposed problem will provide a more ideal solution of F as clustering results. $F^\top \mathbf{1} = 0$ forces each element has an equal chance to be zeros. We use K-means to refine the final results when we obtain the solution F. This non-smooth objective function is very hard to optimize. Thus, in next section, we will introduce an iterative algorithm to solve the problem. Different treatments are utilized for numerator and denominator to ensure that the numerator is reduced while increasing the denominator.

Optimization Algorithm

The $\ell_{2,1}$-norm relaxation of multi-way graph cut clustering is very difficult to optimize because we need to minimize the ratio of non-smooth terms. The existing optimization methods cannot solve this problem effectively. To address this problem, we introduce a new iterative algorithm to get the ideal solution. We change it into a quadratic problem on the Stiefel manifold (QPSM) to optimize.

The first step of optimization is to minimize:

$$\min_{F^\top F=I, F^\top \mathbf{1}=0} \sum_{i,j} W_{ij}\|f_i - f_j\|_2. \tag{13}$$

The re-weighted weight matrix is used to obtain clear cluster structures (Nie et al. 2011). On the basis of similarity matrix W, the re-weighted weight matrix is constructed, defined as:

$$\tilde{W}_{ij} = \frac{W_{ij}}{2\|f_i - f_j\|_2}. \tag{14}$$

Then, we have:

$$\min_{F^\top F=I, F^\top \mathbf{1}=0} \sum_{i,j} \tilde{W}_{ij}\|f_i - f_j\|_2^2. \tag{15}$$

It is very similar to traditional multi-way ratio cut clustering method. We redefine the Laplace matrix $\tilde{L} = \tilde{D} - \tilde{W}$, where $\tilde{D}$ is a diagonal matrix with $\tilde{D}_{ii} = \sum_j \tilde{W}_{ij}$. Then the numerator of Eq. (12) can be formulated as:

$$\min_{F^\top F=I, F^\top \mathbf{1}=0} Tr(F^\top \tilde{L}F). \tag{16}$$

The second step of optimization is to solve the following problem while increasing Eq. (13):

$$\max_{F^\top F=I, F^\top \mathbf{1}=0} \sum_{i,j} \|f_i - f_j\|_2. \tag{17}$$

As suggested in Nie, Yuan, and Huang (2014), we can first solve the following problem:

$$\max_x \sum_i h_i(g_i(x)). \tag{18}$$

If we want to maximize the objective in Eq. (18), we only need to guarantee $h_i(x)$ is a convex function in the range of $g_i(x)$. And then, Eq. (18) can be changed to:

$$\max_x \sum_i Tr(D_i^\top (g_i(x)), \tag{19}$$

where D_i is any subgradient of h_i at point $g_i(x)$. The convergence has been proved in Nie, Yuan, and Huang (2014). In Eq. (17), we let $g_{i,j} = f_i - f_j$ and $h = \|\bullet\|_2$ is an convex function in the domain of $g_{i,j}$. Hence, Eq. (17) can be formulated as:

$$\max_{F^\top F=I, F^\top \mathbf{1}=0} \sum_{i,j} Tr(D_{i,j}^\top (f_i - f_j)), \tag{20}$$

where $D_{i,j} = \frac{f_i - f_j}{\|f_i - f_j\|_2}$, and then it can be further converted into:

$$\max_{F^\top F=I, F^\top \mathbf{1}=0} Tr(F^\top B), \tag{21}$$

where $B = L_c F$ is calculated with current solution F and $L_{c(i,j)} = \sum_j \frac{1}{\|f_i - f_j\|_2} - \frac{1}{\|f_i - f_j\|_2}$.

Based on the above derivation, the optimization method can be translated into the following problem:

$$\min_{F^\top F=I, F^\top \mathbf{1}=0} \frac{Tr(F^\top \widetilde{L}F)}{Tr(F^\top B)}. \tag{22}$$

During each iteration, we calculate $\lambda = \frac{\sum_{i,j} W_{ij}\|f_i - f_j\|_2}{\sum_{i,j}\|f_i - f_j\|_2}$, $\tilde{L}$ and B with the current solution F, and then we can try to update F by minimizing:

$$\min_{F^\top F=I, F^\top \mathbf{1}=0} Tr(F^\top \widetilde{L}F) - \lambda Tr(F^\top B), \tag{23}$$

where the constraint $F^\top \mathbf{1} = 0$ can be changed to:

$$\min_{F^\top F=I} Tr(F^\top (\widetilde{L} + \infty \mathbf{1}\mathbf{1}^\top)F) - \lambda Tr(F^\top B). \tag{24}$$

When we minimize Eq. (24), $Tr(F^\top \mathbf{1}\mathbf{1}^\top F)$ must be zero because of the parameter is infinity. That is to say, $Tr((F^\top \mathbf{1})(F^\top \mathbf{1})^\top) = 0$. Eq. (24) is a quadratic problem on the Stiefel manifold (QPSM). Motivated by Nie, Zhang, and Li (2017), it can be relaxed into:

$$\max_{F^\top F=I} Tr(F^\top \widetilde{A} F) + \lambda Tr(F^\top B), \tag{25}$$

where $\widetilde{A} = \alpha I_m - (\widetilde{L} + \infty \mathbf{1}\mathbf{1}^\top)$ and the relaxation parameter α is an arbitrary constant to guarantee that $\widetilde{L}$ is a positive definite (pd) matrix. Hence, the Lagrangian function for Eq. (25) can be formulated as:

$$L_1(W, \Lambda) = Tr(F^\top \widetilde{A} F) + \lambda Tr(F^\top B) - Tr(\Lambda(F^\top F - I)), \tag{26}$$

the KKT condition for the problem Eq. (26) can be calculated as:

$$\frac{\partial L_1}{\partial F} = 2\widetilde{A}F + \lambda B - 2F\Lambda. \tag{27}$$

The orthogonal iteration is utilized to solve this problem by finding the first k dominant eigenvalues and their associated eigenvectors, we first construct matrix $M = 2\widetilde{A}F + \lambda B$ and then calculate F via the compact QR factorization of M. The following iterative algorithm is used to solve this problem:

Algorithm 1 Algorithm to solve the problem (25)

Initialize F satisfying $F^\top F = I$
while not converge **do**

1: Update $M \leftarrow 2\widetilde{A}F + \lambda B$
2: Calculate $USV^\top = M$ via the campact SVD method of M
3: Update $F \leftarrow UV^\top$

end while

To sum up, the overall framework of our algorithm is:

Algorithm 2 Algorithm to solve the problem (12)

Initialize F satisfying $F^\top F = I$
while not converge **do**

1: Calculate $\lambda = \frac{\sum_{i,j} W_{ij}\|f_i - f_j\|_2}{\sum_{i,j}\|f_i - f_j\|_2}$ and the matrix S, B, where $S_{i,j} = \frac{1}{2\|f_i - f_j\|_2}$, $B = L_c F$ and $L_{c(i,j)} = \sum_j \frac{1}{\|f_i - f_j\|_2} - \frac{1}{\|f_i - f_j\|_2}$
2: Update F with the Eq. (23), where $\tilde{L} = \tilde{D} - \tilde{W}$,$\tilde{W} = W \circ S$ and $\tilde{D}$ is a diagonal matrix with $\tilde{D}_{ii} = \sum_j \tilde{W}_{ij}$

end while

Complexity and Convergence Analysis

Our method is composed of two sub-problems. The complete algorithm is shown in Algorithm 2. The complexity of first step is $O(n^2 + nc)$, c and n is the number of clusters and the number of data. Under the condition $c \ll n$, the total complexity is basically $O(n^2)$, and the complexity of the second step $O(tn^2c)$ (Nie, Zhang, and Li 2017), where t is the number of iteration of the QPSM when we update the solution. So our method doesn't have high computational cost and the total complexity is $O(kn^2 + kn^2ct)$, where k is the number of iterations of Algorithm 2.

The converges of Algorithm 1 has been proved in Nie, Zhang, and Li (2017), and in the following, we will prove that the algorithm 2 will monotonically decrease until convergence.

Proof. According to the Algorithm 1, we can know that:

$$F_{t+1} = \min_{F^\top F=I, F^T 1=0} Tr(F^\top \widetilde{L_t} F) - \lambda Tr(F^\top B_t). \tag{28}$$

That is to say:

$$\begin{aligned} &Tr(F_{t+1}^\top \widetilde{L_t} F_{t+1}) - \lambda Tr(F_{t+1}^\top B_t) \\ &\leq Tr(F^\top \widetilde{L_t} F) - \lambda Tr(F^\top B_t). \end{aligned} \tag{29}$$

According to the property of trace , we have:

$$\begin{aligned} &Tr(F_{t+1}^\top \widetilde{L_t} F_{t+1}) - \lambda Tr(B_t F_{t+1}^\top) \\ &\leq Tr(F^\top \widetilde{L_t} F) - \lambda Tr(B_t F^\top). \end{aligned} \tag{30}$$

From Eq. (30), we can easily know that:

$$\begin{aligned} &\sum_{i,j}^n \frac{W_{ij}\|f_{t+1}^i - f_{t+1}^j\|_2^2}{2\|f_t^i - f_t^j\|_2} - \lambda \sum_{i,j}^n \frac{(f_{t+1}^i - f_{t+1}^j)(f_t^i - f_t^j)^T}{\|f_t^i - f_t^j\|_2} \\ &\leq \sum_{i,j}^n \frac{W_{ij}\|f_t^i - f_t^j\|_2^2}{2\|f_t^i - f_t^j\|_2} - \lambda \sum_{i,j}^n \frac{(f_t^i - f_t^j)(f_t^i - f_t^j)^T}{\|f_t^i - f_t^j\|_2}. \end{aligned} \tag{31}$$

Nie et al. (2011) proved that if we have any nonzero vectors f and f_t, the following inequality holds:

$$\begin{aligned} (\sqrt{f} - \sqrt{f_t})^2 \geq 0 \Rightarrow f - 2\sqrt{f f_t} + f_t \geq 0 \\ \Rightarrow \sqrt{f} - \frac{f}{2\sqrt{f_t}} \leq \sqrt{f_t} - \frac{f_t}{2\sqrt{f_t}}. \end{aligned} \tag{32}$$

According to Eq. (31), we have:

$$\begin{aligned} &\sum_{i,j}^n W_{ij}(\|f_{t+1}^i - f_{t+1}^j\|_2 - \frac{\|f_{t+1}^i - f_{t+1}^j\|_2^2}{2\|f_t^i - f_t^j\|_2}) \\ &\leq \sum_{i,j}^n W_{ij}(\|f_t^i - f_t^j\|_2 - \frac{\|f_t^i - f_t^j\|_2^2}{2\|f_t^i - f_t^j\|_2}). \end{aligned} \tag{33}$$

Adding the above two inequalities in Eq. (31) and (33), we can know that:

$$\begin{aligned} &\sum_{i,j}^n W_{ij}\|f_{t+1}^i - f_{t+1}^j\|_2 - \lambda \sum_{i,j}^n \frac{(f_{t+1}^i - f_{t+1}^j)(f_t^i - f_t^j)^\top}{\|f_t^i - f_t^j\|_2} \\ &\leq \sum_{i,j}^n W_{ij}\|f_t^i - f_t^j\|_2 - \lambda \sum_{i,j}^n \frac{(f_t^i - f_t^j)(f_t^i - f_t^j)^\top}{\|f_t^i - f_t^j\|_2}. \end{aligned} \tag{34}$$

Notice the definition of $\lambda = \frac{\sum_{i,j}^n W_{ij}\|f_t^i - f_t^j\|_2}{\sum_{i,j}^n \|f_t^i - f_t^j\|_2}$, we can easily know that:

$$\sum_{i,j}^n W_{ij}\|f_t^i - f_t^j\|_2 - \lambda \sum_{i,j}^n \frac{(f_t^i - f_t^j)(f_t^i - f_t^j)^\top}{\|f_t^i - f_t^j\|_2} \leq 0. \tag{35}$$

Then we have:

$$\frac{\sum_{i,j}^{n} W_{ij}\|f_{t+1}^{i}-f_{t+1}^{j}\|_2}{\sum_{i,j}^{n}\|f_{t+1}^{i}-f_{t+1}^{j}\|_2} \leq \lambda \frac{\sum_{i,j}^{n} \frac{(f_{t+1}^{i}-f_{t+1}^{j})(f_{t}^{i}-f_{t}^{j})^{\top}}{\|f_{t}^{i}-f_{t}^{j}\|_2}}{\sum_{i,j}^{n}\|f_{t+1}^{i}-f_{t+1}^{j}\|_2} \leq \lambda = \frac{\sum_{i,j}^{n} W_{ij}\|f_{t}^{i}-f_{t}^{j}\|_2}{\sum_{i,j^{n}}\|f_{t}^{i}-f_{t}^{j}\|_2}. \tag{36}$$

Therefore, the algorithm will monotonically decrease the Eq. (12) in each iteration until convergence. □

Experimental Results

In this section, we evaluate the effectiveness of proposed graph clustering method in both synthetic data and real datasets, and then compare the performance with several state-of-art clustering methods.

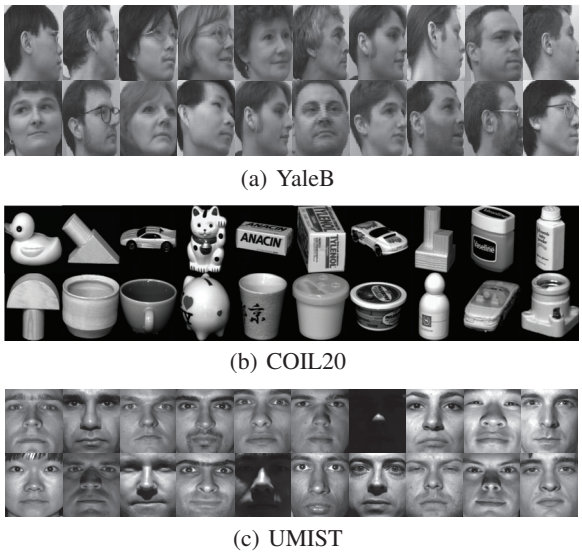
To evaluate the clustering results, we adopt two standard evaluation metrics: Accuracy (ACC) and Normalized Mutual Information (NMI) (Cai et al. 2008). We construct the original weight matrix W with probabilistic K-nearest neighbors for each datasets. The weight W_{ij} is calculated as nearest-neighbor graph (Gu and Zhou 2009), and the number of neighbors is set to 5. We adopt K-means to refine the final results, and repeat K-means for 50 times with random initializations.

Baselines

We compared the performance of our clustering method with Ratio Cut (RCut) (Chan, Schlag, and Zien 1994), Normalized Cut (NCut) (Ng, Jordan, and Weiss 2002), ℓ_1-norm Relaxations for Graph Clustering (NR-RC) (Nie et al. 2016a), which recursively run the clustering method to obtain the desired number of partitions, Large Scale Spectral Clustering (LSC) (Chen and Cai 2011), Self-tuning spectral clustering (RLS) (Zelnik-Manor and Perona 2005), Agglomerative clustering on a directed graph (GAC) (Zhang et al. 2012) and the constrained Laplacian rank (CLR) (Nie et al. 2016b) method, which learn a new data graph with exactly c connected components (where c is the number of clusters). Specifically, CLR denotes the method using ℓ_1-norm distance clustering. Some algorithms are based on K-means and depend on the initialization. Without losing generality, we repeat K-means for 50 times with random initializations, and then we report the results corresponding to the best objective values.

Evaluations in Synthetic Data

In this experiment, we first evaluate the proposed method in the synthetic data, including randomly generated data points distributed on three ring shapes with noise. In Figure 2, we set the color of the three clusters to be red, black and blue, respectively. Figure 2(a) is the original data where different colors represent different clusters. Figure 2(b) shows the results of RCut (Chan, Schlag, and Zien 1994) and Figure 2(c) shows the results of NCut (Ng, Jordan, and Weiss 2002). The last three figures are the results of our method with iterating, where t is the number of iterations. From these figures, we can observe that the solution of RCut and NCut may deviate from the ideal ones, even though these data are not difficult to split, and our algorithm can gradually revise the clustering results in the iterative process.



(a) YaleB

(b) COIL20

(c) UMIST

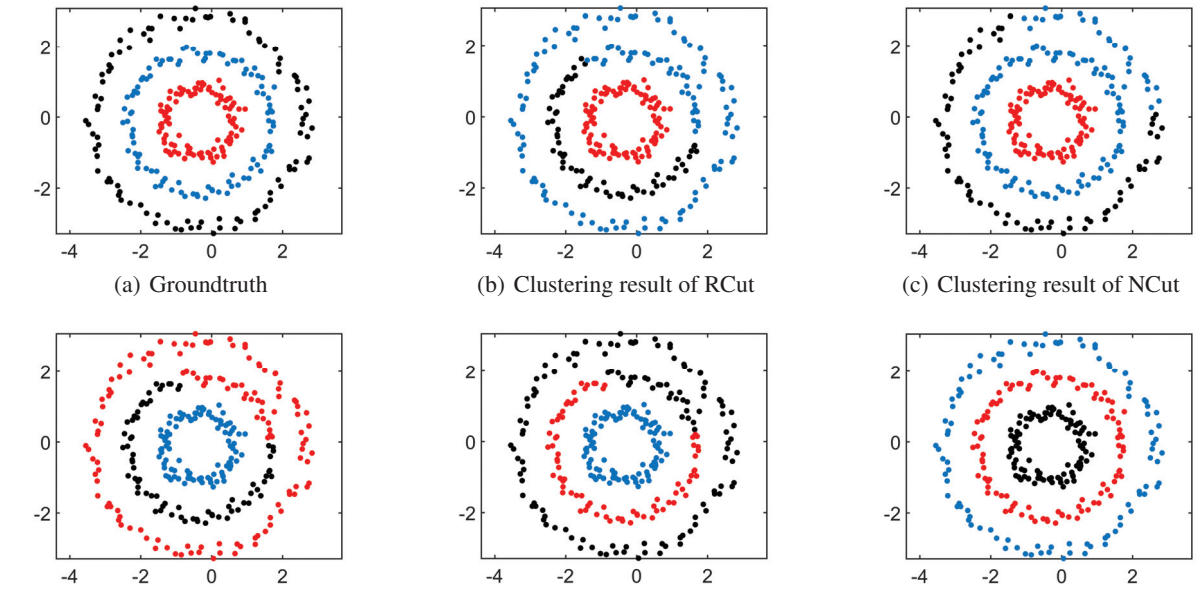
Figure 1: The image samples from the three benchmark datasets used in our experiments

Table 1: Description of Datasets

Datasets	#Size	#Dimensions	#Classes
COIL20	1440	1024	20
Yale	165	1024	15
YaleB	2414	1024	38
COIL100	7200	1024	100
Yeast	1484	1470	10
UMIST	575	1024	20
MSRA	1799	256	12
BINALPHA	1404	320	36
DERMATOL	366	34	6

Evaluations in Real Image Datasets

In this experiment, we evaluate our proposed method on nine benchmark datasets as follows. Two UCI datasets are Dermatol and Yeast (Asuncion and Newman 2007). Two object datasets COIL20 and COIL100 (Nene et al. 1996) contain different objects imaged at every angle in a 360 rotation and the backgrounds have been discarded. One digit and character dataset BINALPHA contains 26 binary hand-written alphabets and 10 numbers. Then the four face datasets are adopted. Yale contains 165 gray-scale images in 15 individuals, one per different facial expression or configuration. YaleB (Georghiades, Belhumeur, and Kriegman 2001) has 38 individuals and around 64 near frontal images under different illuminations per individual. We simply use the cropped images and resize them to 32 $\times$ 32 pixels. Umist (Graham and Allinson 1998) consists of 20 individuals, sub-



(a) Groundtruth (b) Clustering result of RCut (c) Clustering result of NCut

(d) Clustering result of our method, $t = 1$, $ACC = 0.9132$ (e) Clustering result of our method, $t = 3$, $ACC = 0.9232$ (f) Clustering result of our method, $t = 5$, $ACC = 1$

Figure 2: Clustering result in synthetic data

Table 2: Clustering accuracy (%) comparison of multi-way clustering on the nine datasets

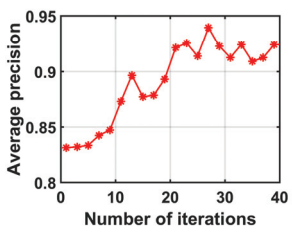
DATA SET	RCut	CLR	LSC	RLS	GCA	NR-RC	Our Method
COIL20	81.64(2.29)	86.39(0.58)	79.82(3.15)	82.50(0.00)	83.26(0.00)	81.32(2.18)	**92.27(2.12)**
Yale	64.84(2.15)	59.39(0.34)	60.09(2.23)	61.87(0.13)	48.48(0.00)	55.76(1.25)	**68.12(0.74)**
YaleB	28.91(0.67)	29.08(0.54)	18.27(0.94)	28.67(0.68)	25.60(0.00)	29.80(0.79)	**42.05(0.96)**
COIL100	70.78(1.73)	76.38(1.71)	58.38(1.16)	69.90(1.19)	74.33(0.00)	71.56(1.77)	**77.31(1.25)**
Yeast	46.05(0.14)	46.67(0.03)	32.68(3.40)	40.23(0.00)	42.92(0.00)	40.66(1.25)	**50.13(2.02)**
UMIST	66.04(1.35)	74.61(1.92)	64.78(2.96)	66.22(2.96)	61.39(0.02)	73.91(3.25)	**80.48(3.08)**
MSRA	60.03(1.18)	60.08(2.11)	57.70(3.26)	56.03(0.00)	45.14(0.00)	61.42(1.78)	**64.15(1.45)**
BINALPHA	47.42(0.98)	45.73(1.10)	47.08(1.76)	47.08(1.12)	48.79(0.00)	46.75(1.02)	**51.85(0.70)**
DERMATOL	77.32(0.00)	77.60(0.00)	84.07(4.21)	77.32(0.00)	84.32(0.00)	78.95(0.82)	**85.79(0.00)**

Table 3: NMI (%) comparison of multi-way clustering on the nine datasets

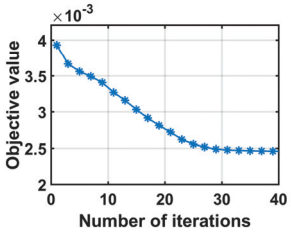
DATA SET	RCut	CLR	LSC	RLS	GCA	NR-RC	Our Method
COIL20	90.54(1.27)	94.50(0.89)	88.55(1.09)	91.02(0.00)	92.29(0.00)	89.64(2.12)	**95.76(1.25)**
Yale	62.02(0.41)	58.94(0.39)	63.86(1.63)	60.25(0.11)	58.69(0.00)	59.67(0.82)	**69.74(1.02)**
YaleB	34.27(0.83)	33.88(0.58)	26.44(0.85)	31.81(0.51)	34.77(0.00)	38.70(0.72)	**52.23(0.43)**
COIL100	87.31(1.31)	91.98(1.32)	81.47(0.69)	88.24(0.47)	91.78(0.00)	90.35(1.39)	**92.18(1.24)**
Yeast	**31.34(0.08)**	30.45(0.08)	21.66(2.17)	27.41(0.00)	27.56(0.00)	23.85(0.47)	30.22(0.20)
UMIST	86.31(0.98)	88.17(1.36)	82.31(1.86)	86.24(2.30)	82.19(4.09)	85.96(1.39)	**89.80(1.75)**
MSRA	70.60(2.17)	70.67(2.42)	72.65(4.47)	66.03(0.76)	49.52(0.00)	75.01(2.08)	**75.94(1.69)**
BINALPHA	63.64(0.68)	62.70(0.63)	60.18(0.81)	61.81(0.37)	64.06(0.00)	61.90(0.52)	**66.11(0.41)**
DERMATOL	86.83(0.00)	**87.75(0.00)**	83.19(2.83)	86.83(0.00)	86.83(0.00)	80.83(0.76)	79.71(0.00)

jects covering a range of race, sex and appearance, and each individual is shown in a range of poses from profile to frontal views. MSRA (Liu et al. 2007) contains 12 individuals with different background and illumination conditions. Some image samples are shown in Figure 1. The brief descriptions of these nine datasets are given in Table 1.

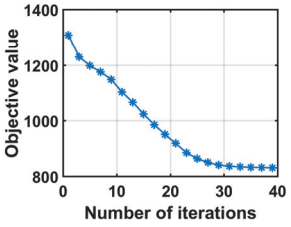
The clustering results are shown in Table 2 and Table 3. We run 20 times for each method and report the mean val-



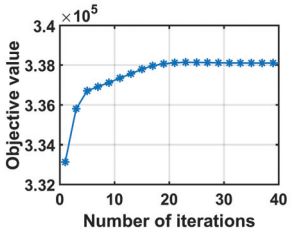
(a) Clustering accuracy $vs.$ iteration on COIL20 dataset



(b) Objective value $vs.$ iteration

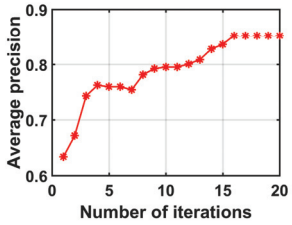


(c) Objective value $vs.$ iteration

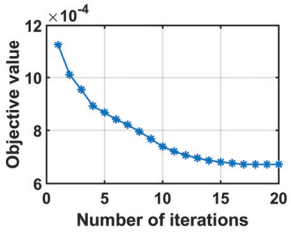


(d) Objective value $vs.$ iteration

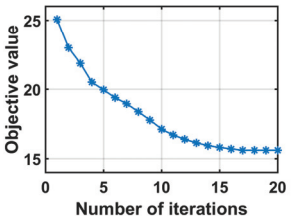
Figure 3: Convergence analysis on COIL20 dataset



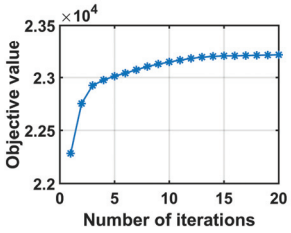
(a) Clustering accuracy $vs.$ iteration on DERMATOL dataset



(b) Objective value $vs.$ iteration



(c) Objective value $vs.$ iteration



(d) Objective value $vs.$ iteration

Figure 4: Convergence analysis on DERMATOL dataset

ues and standard deviations. In these tables, we can observe that our proposed method outperforms the competing methods on these real-world datasets. We find that the proposed method can improve the clustering performance whether in face datasets or in other object datasets. Especially when performing on the object dataset COIL20, the clustering accuracy is over 92%. Note that even though all clustering methods work well on UMIST dataset, our proposed method still outperforms two of the best methods, CLR and RCut. The COIL100 dataset is very difficult to deal with due to the complexity of classes, but our method still harvests a good result. To sum up, the performance of these baselines in different datasets are inconsistent. For instance, LSC works better on Yale and CLR works better on UMIST. Yet the performances of our method increase significantly in all the datasets.

In addition, we need to investigate the convergence properties and some details of our algorithm because our method is an iterative algorithm. We verify the convergence of the algorithm in three datasets, and the change of the clustering accuracy with each iteration. The change of accuracy with each iteration of COIL20 data is shown in Figure 3(a), and DERMATOL data in Figure 4(a). The objective value is computed by Eq. (12) for our proposed method, which are plotted in Figure 3(b) for COIL20 data, and Figure 4(b) for DERMATOL data. In order to better understand some of the details of the algorithm in the iterative process, the change of Eq. (13) and Eq. (17) with each iteration are plotted in Figure 3(c) and Figure 3(d) for COIL20 data, Figure 4(c) and Figure 4(d) for DERMATOL data. From these figures, we can find out that our algorithms can effectively increase the similarity within the cluster, increasing the difference between different clusters, which can obtain a better clustering performance. However, the K-means algorithm converges to local optimum and causes the clustering results fluctuate up and down sometimes while the performance of our method is steady growth on the whole.

Conclusion

This work proposed a novel relaxation for the multi-way ratio graph cut clustering method. Specifically, the $\ell_{2,1}$-norm distances instead of the squared distance are adopted for the normalized clustering solution, which can guarantee the solution to being optimal. That is to say, our method can obtain the discrete and sparse indicator vectors with many zeros to make the cluster structures more clear. Furthermore, to efficiently optimize the ratio of two non-smooth items, we convert the optimization problem into a quadratic problem on the Stiefel manifold (QPSM), which takes much less times for the convergence. Extensive experiments demonstrate our method consistently achieves better clustering results than several traditional clustering methods.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 61572388, Grant 61402026 and Grant 61703327, and in part by the Key R&D Program-The Key Industry Innovation Chain of Shaanxi under Grant 2017ZDCXL-GY-05-04-02 and Grant 2017ZDCXL-GY-05-04-02.

References

An, L.; Gao, X.; Li, X.; Tao, D.; Deng, C.; and Li, J. 2012. Robust reversible watermarking via clustering and enhanced pixel-wise masking. *IEEE Transactions on image processing* 21(8):3598–3611.

Asuncion, A., and Newman, D. 2007. Uci machine learning repository.

Ben-Hur, A.; Horn, D.; Siegelmann, H. T.; and Vapnik, V. 2001. Support vector clustering. *Journal of machine learning research* 2(Dec):125–137.

Cai, D.; He, X.; Wu, X.; and Han, J. 2008. Non-negative matrix factorization on manifold. In *Data Mining, 2008. Eighth IEEE International Conference on ICDM'08*, 63–72. IEEE.

Chan, P. K.; Schlag, M. D.; and Zien, J. Y. 1994. Spectral k-way ratio-cut partitioning and clustering. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 13(9):1088–1096.

Chen, X., and Cai, D. 2011. Large scale spectral clustering with landmark-based representation. In *AAAI*, volume 5, 14.

Cheng, C.-K., and Wei, Y.-C. 1991. An improved two-way partitioning algorithm with stable performance (vlsi). *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 10(12):1502–1511.

Deng, C.; Ji, R.; Liu, W.; Tao, D.; and Gao, X. 2013. Visual reranking through weakly supervised multi-graph learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 2600–2607.

Deng, C.; Ji, R.; Tao, D.; Gao, X.; and Li, X. 2014. Weakly supervised multi-graph learning for robust image reranking. *IEEE transactions on multimedia* 16(3):785–795.

Georghiades, A. S.; Belhumeur, P. N.; and Kriegman, D. J. 2001. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE transactions on pattern analysis and machine intelligence* 23(6):643–660.

Graham, D. B., and Allinson, N. M. 1998. Characterising virtual eigensignatures for general purpose face recognition. In *Face Recognition*. Springer. 446–456.

Grauman, K., and Darrell, T. 2006. Unsupervised learning of categories from sets of partially matching image features. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, 19–25. IEEE.

Gu, Q., and Zhou, J. 2009. Co-clustering on manifolds. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 359–368. ACM.

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

Liu, T.; Sun, J.; Zheng, N.-N.; Tang, X.; and Shum, H.-Y. 2007. Learning to detect a salient object. In *Computer Vision and Pattern Recognition, 2007. IEEE Conference on CVPR'07*, 1–8. IEEE.

Liu, W.; Tsang, I. W.; and Müller, K.-R. 2017. An easy-to-hard learning paradigm for multiple classes and multiple labels. *Journal of Machine Learning Research* 18(94):1–38.

Nene, S. A.; Nayar, S. K.; Murase, H.; et al. 1996. Columbia object image library (coil-20).

Ng, A. Y.; Jordan, M. I.; and Weiss, Y. 2002. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, 849–856.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. H. 2010. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *Advances in neural information processing systems*, 1813–1821.

Nie, F.; Wang, H.; Huang, H.; and Ding, C. 2011. Unsupervised and semi-supervised learning via ℓ_1-norm graph. In *IEEE International Conference on Computer Vision (ICCV)*, 2268–2273. IEEE.

Nie, F.; Wang, H.; Deng, C.; Gao, X.; Li, X.; Huang, H.; et al. 2016a. New l1-norm relaxations and optimizations for graph clustering. In *AAAI*, 1962–1968.

Nie, F.; Wang, X.; Jordan, M. I.; and Huang, H. 2016b. The constrained laplacian rank algorithm for graph-based clustering. In *AAAI*, 1969–1976.

Nie, F.; Yuan, J.; and Huang, H. 2014. Optimal mean robust principal component analysis. In *International Conference on Machine Learning*, 1062–1070.

Nie, F.; Zhang, R.; and Li, X. 2017. A generalized power iteration method for solving quadratic problem on the stiefel manifold. *Science China Information Sciences* 60(11):112101.

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

Xu, W.; Liu, X.; and Gong, Y. 2003. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 267–273. ACM.

Zelnik-Manor, L., and Perona, P. 2005. Self-tuning spectral clustering. In *Advances in neural information processing systems*, 1601–1608.

Zhang, W.; Wang, X.; Zhao, D.; and Tang, X. 2012. Graph degree linkage: Agglomerative clustering on a directed graph. *Computer Vision–ECCV 2012* 428–441.