DIRECTLY SOLVING THE ORIGINAL RATIOCUT PROBLEM FOR EFFECTIVE DATA CLUSTERING

Jing Li, Feiping Nie, Xuelong Li

School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, P. R. China

ABSTRACT

This paper focuses on the original RatioCut problem, which is one of the most representative clustering paradigms. The RatioCut criterion looks for a partition of the graph to achieve the mincut cost while keeping each partition reasonably large. This well-known problem is NP hard and its relaxed form has been widely used in the past several decades. However, the relaxed RatioCut usually suffers two problems: not satisfactory stable clustering performance, and undesired two-stage optimization. In this work, we solve the original RatioCut problem by learning a new similarity matrix which has as many connected components as the cluster number, so that the original RatioCut constraint can be directly satisfied. An easily implemented algorithm is derived to iteratively optimize the proposed method. Experimental results on various real-world benchmark datasets exhibit the effectiveness of the proposed method to solve the RatioCut problem.

***Index Terms*—** RatioCut clustering, discrete constraint conditions, iterative optimization

1. INTRODUCTION

Graph-based clustering methods have been successful in data analysis. In the context of graph cut, one needs to find a partition of the graph such that the edges between different groups are exactly cut off. Since the edges between different groups have the low weights while they have the high weights within the same group, the mincut cost is often used as the objective. A very early work about the bipartition to a graph has been discussed in [1]. To prevent the solution of mincut simply separating one individual vertex from the rest of the graph, RatioCut [2] and Ncut [3] explicitly request that every cluster is reasonably large. However, such a balancing condition makes them become NP hard [4], thus many subsequent popular works trickly use eigenvalue decomposition to solve a relaxed problem.

Although many recent applications have successfully employed graph cut technique in their domains, such as graph clustering [5, 6], image segmentation [7, 8], and saliency detection [9], few of them concern about a feasible solution to the original graph cut problem. Specifically, most of these methods first adopt an optimization strategy to the given graph to get a relaxed continuous solution and then discretize it with K-means or spectral rotation algorithms (see [10] and discussion therein). One disadvantage of this kind of approaches is that the final clustering structures are not identical with the data graph (Note that postprocessing like K-means algorithm itself is not a stable operation). The similar consideration is also taken into by [11], where they tackle this problem by the graph approximation. Actually, for the graph-cut problem, a welcome expectation is that we can directly learn the discrete solution which strictly satisfies the discrete constraint conditions.

In this work we propose a novel graph-based clustering method, which can be used to effectively solve the original RatioCut problem. To be specific, we introduce a new similarity matrix, which is supposed to have c (c is the cluster number) connected components, and helps suppress the constraints to be satisfied. Hence, there is no need to switch between discrete and continuity for the target variable in our method. Different from the previous methods (e.g., unnormalized spectral clustering first calculates the eigenvectors of Laplacian matrix and then clusters the derived representation of each data with K-means algorithm), we solve the proposed objective via an easily implemented iterative optimization algorithm. To validate the effectiveness of the proposed method, empirical studies are conducted on different real-world benchmark data sets. The experimental results demonstrate the proposed method outperforms other compared methods in most cases, and more importantly it significantly improves the performance of the relaxed RatioCut.

Notation: Throughout the paper, every matrix is written as uppercase. For a matrix M, the i-th row, the j-th column, and the ij-th element of M are denoted by m_i, m^j, and m_{ij}, respectively. The trace of matrix M is denoted by $Tr(M)$. The L2-norm of vector v is denoted by $\|v\|_2$, and the Frobenius norm of matrix M is denoted by $\|M\|_F$.

2. PROBLEM FORMULATION

Given an undirected data graph described by a similarity matrix $W \in \mathbb{R}^{n\times n}$, where n is the number of data points, the main task of clustering is to partition these points into c group-

s. We use $Y \in \mathbb{R}^{n\times c}$ to denote the indicator matrix, where $y_{ij} = 1$ if the i-th data is assigned to the j-th cluster, it is 0 otherwise. Let $Y = \left[y^1, y^2, ..., y^c\right]$, and then c-way RatioCut criterion is described as an optimization program of variable Y [2]:

$$\max_{Y} \frac{1}{c}\sum_{j=1}^{c}\frac{{y^j}^T W y^j}{{y^j}^T y^j} \quad s.t. Y \in \{0,1\}^{n\times c}, Y1_c = 1_n, \quad (1)$$

where 1_d denotes a d dimensional column vector of all 1's. By defining a scaled cluster assignment matrix $F \in \mathbb{R}^{n\times c}$ as

$$F = Y\left(Y^T Y\right)^{-\frac{1}{2}}, \quad (2)$$

the problem (1) is easily verified to be equivalent to the following form (where the const $\frac{1}{c}$ has been omitted)

$$\min_{F\in Disc} Tr\left(F^T L F\right), \quad (3)$$

which is compact and has been widely used in previous clustering works [12, 13]. In problem (3), L is the so-called Laplacian matrix and $L = D - W$, where the degree matrix $D \in \mathbb{R}^{n\times n}$ is a diagonal matrix whose i-th diagonal element is $\sum_j w_{ij}$, $Disc$ is short for *discrete* and this constraint represents that F should satisfy the Eq. (2), where Y is discrete as constrained in the problem (1). Obviously, the problem (3) is NP hard and the optimal discrete solution cannot be directly obtained. Thus, as stated above, many previous works resort to solving the relaxed problem (i.e., the constraint is reduced to $F^T F = I_c$.) and add a discretizing postprocessing, which makes it a two-stage procedure and not robust in practice.

Intuitively, if the data can be partitioned into c clusters, an ideal neighbors assignment is supposed to contain exactly c connected components. In this paper, we leverage this property by introducing a new similarity matrix $S \in \mathbb{R}^{n\times n}$ with c connected components to help us tackle the problem (3). Since this property doesn't go straightforward with the problem (3), we do the following transformation.

The Laplacian matrix of S is defined as $L_S = D_S - \frac{S^T+S}{2}$, where the degree matrix $D_S \in \mathbb{R}^{n\times n}$ is a diagonal matrix whose i-th diagonal element is $\sum_j (s_{ij} + s_{ji})/2$. L_S is obviously symmetric and positive semi-definite which has the following property [14]:

Theorem 1. *The multiplicity c of the eigenvalue 0 of the Laplacian matrix L_S is equal to the number of connected components in the graph with the similarity matrix S.*

Furthermore, let $\sigma_i(L_s)$ denote the i-th smallest eigenvalue of L_s. According to the Ky Fans Theorem [15], the following equation holds

$$\sum_{i=1}^{c}\sigma_i(L_s) = \min_{F^T F = I_c} Tr(F^T L_S F). \quad (4)$$

Therefore, we can add the right of Eq. (4) to the problem (3), where each F shares the orthogonal constraint. Since we include S as a new variable, to avert that S is too sparse to be capable of forming solid components, a regularized term $\|S\|_F^2$ is introduced. Thus we come to the new objective:

$$\begin{aligned}&\min_{F,S} Tr\left(F^T L F\right) + \alpha\|S\|_F^2 + \lambda Tr(F^T L_S F)\\ &s.t. F^T F = I_c, s_{ij} \geq 0, \sum\nolimits_j s_{ij} = 1,\end{aligned} \quad (5)$$

where α is a positive hyperparameter, each element in S is constrained as nonnegative, and each row of S sums to 1. Particularly, when λ is large enough, according to Eq. (4), we know that the optimal solution S to the problem (5) will make the term $\sum_{i=1}^{c}\sigma_i(L_s)$ equal to zero and thus the similarity matrix S will have c connected components. Under this condition, from Theorem 2 [13] the solution F to the problem (5) is consistent with the *Disc* constraint.

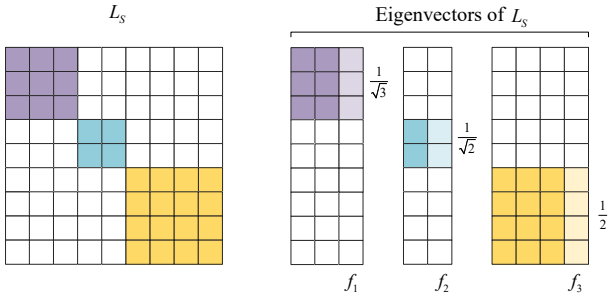

Fig. 1. Left: the learned Laplacian matrix $L_S \in \mathbb{R}^{9\times 9}$ with 3 connected components; Right: the corresponding eigenvectors and the selected components of $F \in \mathbb{R}^{9\times 3}$.

Theorem 2. *If the Laplacian matrix L_S has as many eigenvalues 0 as there are connected components, then the corresponding eigenvectors are the indicator vectors of the connected components.*

Let us take an example. Suppose we have the Laplacian matrix L_S with 3 connected components like Fig. 1 (left) (For better presentation, we rearrange S to be diagonal). Instead of directly operating on L_S, we calculate the eigenvectors of its every diagonal block and then fill in the remaining empty positions with zeros (see Fig. 1 (right)). By arranging the eigenvectors of each block based on the corresponding eigenvalues from large to small, the right-most eigenvector of each diagonal block is $\frac{1}{\sqrt{n_i}}1_{n_i}$ (n_i is the size of a cluster), which corresponds to the eigenvalue 0. Therefore, the scaled cluster assignment matrix $F = [f^1, f^2, f^3]$, which can be verified to exactly satisfy the *Disc* constraint.

3. OPTIMIZATION

Optimizing the problem (5) is still challenging because it involves two variables F, S, and simultaneously they are coupled with each other. In this section, we solve this problem by alternatingly optimizing F and S.

When S is fixed, the problem (5) becomes

$$\begin{aligned}&\min_{F^TF=I} Tr(F^TLF) + \lambda Tr(F^TL_SF)\\ =&\min_{F^TF=I} Tr\left(F^T\left(L+\lambda L_s\right)F\right).\end{aligned} \tag{6}$$

The optimal solution of F is formed by the c eigenvectors of $L+\lambda L_S$ corresponding to the c smallest eigenvalues.

When F is fixed, the problem (5) becomes

$$\min_{s_{ij}\geq 0, \sum_j s_{ij}=1} \alpha \left\|S\right\|_F^2 + \lambda Tr(F^TL_SF). \tag{7}$$

Let $F=[f_1; f_2; ...; f_n]$, then the following equation holds:

$$2Tr(F^TL_SF) = \sum_{i,j} \left\|f_i - f_j\right\|_2^2 s_{ij}. \tag{8}$$

Taking Eq. (8) into the problem (7), we have the following equivalent form

$$\min_{s_{ij}\geq 0, \sum_j s_{ij}=1} \alpha \sum_{i,j} s_{ij}^2 + \frac{\lambda}{2}\sum_{i,j}\left\|f_i - f_j\right\|_2^2 s_{ij}. \tag{9}$$

Considering that the problem (9) is independent for different i, we solve the following problem separately for each i:

$$\min_{s_{ij}\geq 0, \sum_j s_{ij}=1} \alpha \sum_{j} s_{ij}^2 + \frac{\lambda}{2}\sum_{j}\left\|f_i - f_j\right\|_2^2 s_{ij}. \tag{10}$$

Denoting $v_{ij} = \left\|f_i - f_j\right\|_2^2$, and denoting v_i as a vector whose j-th element equals to v_{ij} (and similarly for s_i), the problem (10) can be written in vector form as

$$\min_{s_i\geq 0, s_i 1_n=1} \left\|s_i + \frac{\lambda}{4\alpha}v_i\right\|_2^2. \tag{11}$$

This problem can be solved by [16] or an efficient iterative algorithm proposed in [17].

Based on the above analysis, the detailed procedure for solving the problem (5) is summarized in Algorithm 1.

Algorithm 1 The algorithm to solve the problem (5)

Input: $W\in\mathbb{R}^{n\times n}$, cluster number c, a large enough λ

Initialize $F\in\mathbb{R}^{n\times c}$, which is formed by the c eigenvectors of $L_S = D_S - \frac{W^T+W}{2}$ corresponding to the c smallest eigenvalues

repeat

1. For each i, update the i-th row of S by solving the problem (11)
2. Update F, which is formed by the c eigenvectors of $L+\lambda L_S$ corresponding to the c smallest eigenvalues

until converge

Output: The cluster assignment matrix $F\in\mathbb{R}^{n\times c}$, and the new similarity matrix $S\in\mathbb{R}^{n\times n}$

Table 1. Statistics of seven benchmark datasets.

data set	# of size	# of dimensionality	# of class
Vehicle	846	18	4
Yeast	1484	8	10
Abalone	4177	8	29
Dermatology	1440	1024	20
COIL20	1440	1024	20
USPS	2007	256	10
Umist	575	1024	20

4. ANALYSIS OF ALGORITHM

In this section, we present more insights about the proposed algorithm.

1) *Computing the final clustering result*. When Algorithm 1 converges, we obtain a new similarity matrix S with c connected components, and a neat scaled cluster assignment indicator F. Both of them can be naturally used to partition the original data points into c groups and they obtain the identical results when λ is large enough.

2) *Convergence analysis*. The problem (5) can be divided into two subproblems and we can obtain the optimal solution to each of them. Therefore, by solving the subproblems alternatively, the proposed algorithm will converge to a local solution.

3) *The parameters*. At the first glance, there are two hyperparameters α, λ in the objective of problem (5) which need to be tuned, but actually they can be easily handled. Since α only appears in the subproblem (11), it can be absorbed in λ (Throughout our experiments, we simply fix α to be 1). Particularly, we determine the best λ in a heuristic way to accelerate the computation. Specifically, if there are less than c components in L_S, we multiply λ by two; if more, we divide λ by two; otherwise we stop the iteration. This strategy can effectively relieve the cost for finding the optimal λ.

4) *Time complexity.* The major computation cost in each iteration involves eigenvalue decomposition (solving the subproblem (6)) which is $\mathcal{O}(n^3)$, while optimizing the subproblem (11) is much light, only $\mathcal{O}(n)$. Suppose the needed iterations of the proposed algorithm is T, we conclude that the total complexity is about $\mathcal{O}(Tn^3)$. Comparing with the standard relaxed RatioCut ($\mathcal{O}(n^3+ncdi)$, d is the data dimension, and i is the iterations), our method is expected to perform well at the expense of the larger computation.

5. EXPERIMENTS

In this section, to explore the performance of the proposed method for solving the original RatioCut problem, we conduct our experiments on seven benchmark datasets: Vehicle, Yeast, Abalone, Dermatology (*Dermato*), COIL20, USPS, and Umist, which are briefly summarized in Table 1.

Table 2. Experimental results on seven benchmark datasets.

		Vehicle	Yeast	Abalone	Dermato	COIL20	USPS	Umist
ACC	K-means	0.4421	0.3753	0.1391	0.7514	0.6944	0.6521	0.4452
	RRC	0.4456	0.4111	0.1386	0.9536	0.7813	0.6312	0.4313
	RNC	0.4456	0.3753	0.1515	0.9536	0.7708	0.6358	0.4661
	NMF	0.3995	0.3740	0.1585	0.9508	0.7833	0.6746	**0.4887**
	DRC	**0.4492**	**0.4683**	**0.1764**	**0.9563**	**0.8271**	**0.7165**	0.4870
		vehicle	Yeast	Abalone	Dermato	Coil20	USPS	Umist
NMI	K-means	0.1800	0.2425	0.1542	0.8616	0.7937	0.6299	0.6735
	RRC	0.2131	0.2652	0.1470	0.9051	0.8326	0.7355	0.6766
	RNC	0.2131	0.2401	0.1447	0.9037	0.8387	0.7330	0.6974
	NMF	0.1676	0.2655	0.1460	0.9010	0.8540	0.7568	0.6979
	DRC	**0.2168**	**0.2994**	**0.1565**	**0.9098**	**0.8841**	**0.7718**	**0.7092**

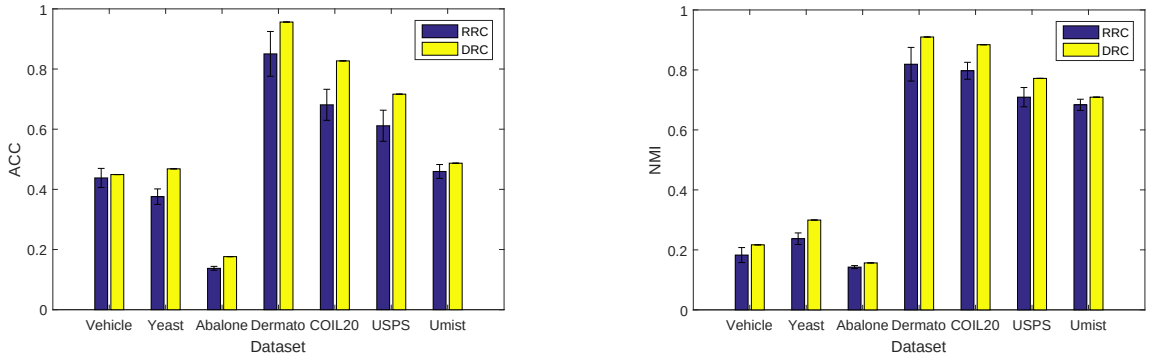

Fig. 2. Clustering results comparison between RRC and the proposed method DRC on seven benchmark datasets

As a convention in [18, 11], the proposed Directly solving RatioCut (DRC) method is compared with K-means, Relaxed RatioCut (RRC), Relaxed Normalized Cut (RNC), NMF [19] methods. The input graph is constructed using the technique proposed in [20], where the Gaussian kernel can be self-tuned (The default number of nearest neighbors is 10 for every data set). For all the methods involving K-means, including K-means itself, we use the random initialization strategy and repeat each of them for 100 times. Their respective results in terms of the minimum K-means value will be reported. As for our method, we run it only once with the initialization described in Algorithm 1, where the maximum number of iterations is set as 30, because we find that our algorithm always converge in 30 times. To measure the final clustering performance, we adopt two regular criteria: the clustering accuracy (ACC) and normalized mutual information (NMI). Table 2 shows the clustering results of each method, where the best result is marked in bold face.

From Table 2, we conclude that the proposed method DRC outperforms the competing methods almost on every data set (Particularly, it shows the great improvement on Yeast, COIL20, and USPS.). The distinct difference among DRC and other graph-based methods is that DRC learns a new similarity matrix after the optimization and thus it is insensitive to the quality of input similarity matrix W. Since W has trifling influence in subproblem (6) when λ is large enough, in this perspective, the input graph can be considered as an initialization of our algorithm.

Since RRC is a representative method for solving the RatioCut problem, it is an important baseline in our experiments. We compute the mean and the corresponding variance of RRC on each dataset over 100 times. The comparison between RRC and DRC is presented as in Fig. 2. It is observed that our once-run DRC impoves RRC both in ACC and NMI. Seeing that RRC obtains very unstable results (noticeable variances) in most datasets, we conclude that the proposed method provides a practical solver for the RatioCut problem.

6. CONCLUSIONS

In this paper, we directly solve the original RatioCut problem. By learning a new similarity matrix with as many connected components as its cluster number, we circumvent the strict discrete constraints in RatioCut and solve the new objective iteratively. The experimental results on seven benchmark datasets prove the effectiveness of the proposed method.

7. REFERENCES

[1] Mechthild Stoer and Frank Wagner, "A simple min-cut algorithm," *Journal of the ACM (JACM)*, vol. 44, no. 4, pp. 585–591, 1997.

[2] 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*, vol. 11, no. 9, pp. 1074–1085, 1992.

[3] Jianbo Shi and Jitendra Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[4] Dorothea Wagner and Frank Wagner, "Between min cut and graph bisection," in *International Symposium on Mathematical Foundations of Computer Science*. Springer, 1993, pp. 744–750.

[5] Linli Xu, Wenye Li, and Dale Schuurmans, "Fast normalized cut with linear constraints," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 2866–2873.

[6] Fei Tian, Bin Gao, Qing Cui, Enhong Chen, and Tie-Yan Liu, "Learning deep representations for graph clustering.," in *AAAI*, 2014, pp. 1293–1299.

[7] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik, "Multiscale combinatorial grouping," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 328–335.

[8] Yipeng Liu, Jing Jin, Qiang Wang, Yi Shen, and Xiaoqiu Dong, "Region level based multi-focus image fusion using quaternion wavelet and normalized cut," *Signal Processing*, vol. 97, pp. 9–30, 2014.

[9] Keren Fu, Chen Gong, Irene Yu-Hua Gu, and Jie Yang, "Normalized cut-based saliency detection by adaptive multi-level region merging," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5671–5683, 2015.

[10] 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*, vol. 22, no. 11, pp. 1796–1808, 2011.

[11] Feiping Nie, Xiaoqian Wang, Michael I Jordan, and Heng Huang, "The constrained laplacian rank algorithm for graph-based clustering," in *Thirtieth AAAI Conference on Artificial Intelligence*. Citeseer, 2016.

[12] Pak K. Chan, Martine D. F. Schlag, and Jason Y. Zien, "Spectral k-way ratio-cut partitioning and clustering," *IEEE Trans. on CAD of Integrated Circuits and Systems*, vol. 13, no. 9, pp. 1088–1096, 1994.

[13] Ulrike Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

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

[15] Ky Fan, "On a theorem of weyl concerning eigenvalues of linear transformations i," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 35, no. 11, pp. 652, 1949.

[16] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra, "Efficient projections onto the l 1-ball for learning in high dimensions," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 272–279.

[17] Jin Huang, Feiping Nie, and Heng Huang, "A new simplex sparse learning model to measure data similarity for clustering," in *Proceedings of the 24th International Conference on Artificial Intelligence*. AAAI Press, 2015, pp. 3569–3575.

[18] Feiping Nie, Xiaoqian Wang, and Heng Huang, "Clustering and projected clustering with adaptive neighbors," in *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, 2014, pp. 977–986.

[19] Feiping Nie, Chris Ding, Dijun Luo, and Heng Huang, "Improved minmax cut graph clustering with nonnegative relaxation," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2010, pp. 451–466.

[20] Lihi Zelnik-Manor and Pietro Perona, "Self-tuning spectral clustering," in *Advances in Neural Information Processing Systems*, 2004, pp. 1601–1608.