Equivalence between Graph Spectral Clustering and Column Subset Selection (Student Abstract)

Guihong Wan[1,2], Wei Mao[3], Yevgeniy R. Semenov[1], Haim Schweitzer[3]

[1]Department of Dermatology, Massachusetts General Hospital, Harvard Medical School, MA, USA
[2]Departments of Biostatistics and Epidemiology, Harvard T. H. Chan School of Public Health, MA, USA
[3]Department of Computer Science, University of Texas at Dallas, Texas, USA
guihong_wan@hsph.harvard.edu, wei.mao@utdallas.edu, ysemenov@mgh.harvard.edu, haim@utdallas.edu

Abstract

The common criteria for evaluating spectral clustering are NCut and RatioCut. The seemingly unrelated column subset selection (CSS) problem aims to compute a column subset that linearly approximates the entire matrix. A common criterion is the approximation error in the Frobenius norm (ApproxErr). We show that any algorithm for CSS can be viewed as a clustering algorithm that minimizes NCut by applying it to a matrix formed from graph edges. Conversely, any clustering algorithm can be seen as identifying a column subset from that matrix. In both cases, ApproxErr and NCut have the same value. Analogous results hold for RatioCut with a slightly different matrix. Therefore, established results for CSS can be mapped to spectral clustering. We use this to obtain new clustering algorithms, including an optimal one that is similar to A*. This is the first nontrivial clustering algorithm with such an optimality guarantee. A variant of the weighted A* runs much faster and provides bounds on the accuracy. Finally, we use the results from spectral clustering to prove the NP-hardness of CSS from sparse matrices.

Introduction

The Graph Spectral Clustering (GSC) Problem

Let G be an undirected graph of n nodes and m edges. Let $w_{ij}{\geq}0$ be the weight between nodes i, j, so that $W{=}(w_{ij})$ is $n{\times}n$. Let $d_i{=}\sum_j w_{ij}$ and $d{=}(d_1,\ldots,d_n)^T$. The $n{\times}n$ degree matrix is $D{=}\mathsf{diag}(d)$. The goal of graph spectral clustering is to compute a partition of the nodes into k disjoint subsets: $A{=}\{A_1,\ldots,A_k\}$. For $t{=}1,\ldots,k$, define: $V_t{=}\sum_{i\in A_t} d_i$, $C_t{=}\sum_{i\in A_t, j\notin A_t} w_{ij}$. Here V_t is the cluster volume, and C_t is the cluster cut. The following criteria are most common for spectral clustering (Von Luxburg 2007):

$$\text{NCut}(A)=\frac{1}{2}\sum_{t=1}^{k}\frac{C_t}{V_t},\ \text{RatioCut}(A)=\frac{1}{2}\sum_{t=1}^{k}\frac{C_t}{|A_t|}. \quad (1)$$

Define the edge vector e associated with nodes i and j as the n-vector: $e{=}\sqrt{w_{ij}}\,(0,\ldots,0,-1,0,\ldots,0,1,0,\ldots,0)^T$, where -1 is in the ith location, and 1 is in the jth location. The vertex-edge matrix and the Laplacian matrix are:

$$E{=}(e_1,e_2,\ldots,e_m),\ L{=}D{-}W{=}EE^T, \quad \text{where } e_i \text{ is the } i\text{th edge vector.} \quad (2)$$

The *normalized* edge vector $\tilde{e}$ associated with nodes i,j is defined as: $\tilde{e}{=}D^{-\frac{1}{2}}e$. The *normalized* vertex-edge matrix and the *normalized* Laplacian matrix are:

$$\tilde{E}{=}(\tilde{e}_1,\tilde{e}_2,\ldots,\tilde{e}_m){=}D^{-\frac{1}{2}}E,\quad \tilde{L}{=}D^{-\frac{1}{2}}LD^{-\frac{1}{2}}{=}\tilde{E}\tilde{E}^T. \quad (3)$$

The Column Subset Selection (CSS) Problem

Let $X{=}(x_1,\ldots,x_m)$ be an $n{\times}m$ matrix. Let X_S be an $n{\times}k$ matrix constructed from k selected columns of X. The linear approximation of X by X_S can be written as: $X\approx X_SB$, where B is a coefficient matrix of size $k\times m$. The squared Frobenius norm of the approximation error is given by:

$$\text{ApproxErr}(X,X_S)=\frac{1}{2}\min_B\|X-X_SB\|_F^2. \quad (4)$$

The goal of the CSS problem is to find X_S from X such that the error criterion (4) is minimized. See e.g., (Shitov 2021; Wan and Schweitzer 2021) for additional details.

Our main contribution is showing the equivalence between graph spectral clustering (GSC), seen here as a clustering method minimizing RatioCut or NCut in (1), and column subset selection (CSS), viewed here as a method to select a column subset minimizing ApproxErr in (4).

Specifically, let G be a graph of n nodes and m edges. We construct a vertex-edge matrix E in (2), or $\tilde{E}$ in (3), of size $n{\times}m$ from G, which has the following property: Any selection of $n{-}k$ linearly independent columns from E or $\tilde{E}$ gives a partition of nodes in G into k clusters. The ApproxErr of CSS is exactly the NCut or RatioCut of GSC. Conversely, clustering the nodes of G into k components identifies a column subset of E or $\tilde{E}$ with rank $n{-}k$.

This relationship implies that established algorithms and theoretical results in the field of CSS can be applied to GSC, and vice versa. Two applications of this are described: (1) We obtain the first nontrivial optimal GSC algorithm from CSS; (2) We prove the NP-hardness of CSS from sparse matrices using the NP-hardness of the NCut problem. See Mao, Wan, and Schweitzer (2024) for additional applications.

Equivalence between GSC and CSS

The proofs of equivalence with RatioCut and NCut are very similar. We give a proof for a generalization that has both of them as special cases.

Algorithm 1: The GSC algorithm from a CSS algorithm

Input: G: a graph. k: the desired number of clusters. Criterion: NCut or RatioCut. A black box CSS algorithm.
Output: k clusters of G.
1. If the criterion is RatioCut, compute E from G. If the criterion is NCut, compute $\tilde{E}$ from G.
2. Run the CSS algorithm to select E_S containing $n-k$ linearly independent columns from E or $\tilde{E}$.
3. Construct the subgraph G_S from the selection.
4. Return the connected components of G_S as output.

We consider a generalized normalization function that associates an arbitrary value $p_i \geq 0$ with each node i. When $p_i=0$, then $d_i=0$ (the node is not connected to other nodes). The results for RatioCut are with $p_i=1$, and for NCut with $p_i=d_i$. Other options are not discussed here.

Set $p=(p_1,\ldots,p_n)^T$. Define $P=\mathsf{diag}(p)$ of size $n\times n$, and $P^{-\frac{1}{2}}$ as $(P^+)^{\frac{1}{2}}$, where P^+ is the pseudoinverse of P. Given a cluster A_t, its volume and cut are: $V_t=\sum_{i\in A_t} p_i$, $C_t=\sum_{i\in A_t, j\notin A_t} w_{ij}$. For a graph partitioned into k clusters: $A=\{A_1,\ldots,A_k\}$, the generalized criterion of RatioCut and NCut in (1) for spectral clustering to minimize is:

$$\mathsf{gCut}(A) = \frac{1}{2}\sum_{t=1}^{k}\frac{C_t}{V_t}. \tag{5}$$

When $p_i=1$, gCut is the same as RatioCut, and when $p_i=d_i$, gCut is the same as NCut. Similarly, we can generalize the notations for e and $\tilde{e}$, E and $\tilde{E}$, L and $\tilde{L}$ in (2) and (3):

$$\tilde{e}=P^{-\frac{1}{2}}e, \quad \tilde{E}=P^{-\frac{1}{2}}E, \quad \tilde{L}=P^{-\frac{1}{2}}LP^{-\frac{1}{2}} = \tilde{E}\tilde{E}^T.$$

Theorem 1. Let G be a graph with the corresponding vertex-edge matrix E. Let S be a subset of edges in G, and let E_S be the corresponding vertex-edge matrix. Without loss of generality, let $n-k$ be the rank of E_S. Let Q_S be an orthonormal basis of $\tilde{E}_S=P^{-\frac{1}{2}}E_S$. Let Q_S^c be the orthogonal complement of Q_S. Define subsets S_1 and S_2 by:

$$S_1 = \{\text{for edges in } G : \tilde{e}^T(Q_S^c)(Q_S^c)^T\tilde{e} = 0\},$$
$$S_2 = \text{all edges not in } S_1.$$

Let E_1 and E_2 be the vertex-edge matrices of S_1 and S_2 respectively. Let $\tilde{E}_2 = P^{-\frac{1}{2}}E_2$ and $\tilde{L}_2 = \tilde{E}_2\tilde{E}_2^T$. Set:

$$\gamma = \frac{1}{2}\mathsf{Trace}[(Q_S^c)^T\tilde{L}_2(Q_S^c)].$$

Then:
(a). The edges in S form a subgraph G_S with k connected components $A=\{A_1,\ldots,A_k\}$.
(b). The edges in S_1 are internal to the clusters $A_1,\ldots,A_k$. The edges in S_2 are between these clusters (cut edges).
(c). $\mathsf{gCut}(A)=\gamma$.
(d). $\text{ApproxErr}(\tilde{E},\tilde{E}_S)=\gamma$.

See the full paper for the proof.

k		$\epsilon=0$	$\epsilon=0.1$	$\epsilon=0.5$	$\epsilon=1$	NJW
2	NCut	**0.069**	0.084	0.084	0.084	**0.069**
	b	**0**	0.063	0.063	0.063	–
3	NCut	**0.147**	**0.147**	0.174	0.174	**0.147**
	b	**0**	0.102	0.129	0.129	–
4	NCut	**0.322**	**0.322**	0.334	0.334	0.330
	b	**0**	0.189	0.200	0.200	–
5	NCut	**0.600**	**0.600**	0.610	**0.600**	0.6625
	b	**0**	0.290	0.3000	0.290	–
6	NCut	**0.878**	**0.878**	**0.878**	**0.878**	0.933
	b	**0**	0.3700	0.3700	0.3700	–

Table 1: Accuracy and bound (b) for NCut: n=20, m=23.

Optimal and Suboptimal Spectral Clustering

As a result of Theorem 1, graph spectral clustering can be solved by column subset selection algorithms (Algorithm 1).

An algorithm, similar to the (weighted) A*, computes optimal and suboptimal solutions for the CSS problem (He et al. 2019). The behavior is controlled by a parameter ϵ. For $\epsilon=0$ the algorithm is optimal but very slow. For $\epsilon>0$ it is faster, not optimal, but gives a bound on how far the computed solution to the optimum.

Running this CSS algorithm as the black box algorihtm within Algorithm 1 gives an optimal clustering algorithm for $\epsilon=0$ and a suboptimal clustering algorithm with guaranteed bounds on accuracy for $\epsilon>0$. Table 1 presents the results of some experiments compared to the classical spectral clustering algorithm (NJW) (Ng, Jordan, and Weiss 2001). The "cockroach" graph was used (Guattery and Miller 1998).

NP-Hardness of Column Subset Selection

The CSS problem was recently proved NP-hard for dense matrices (Shitov 2021). The proof reduces the *three coloring problem* (NP-complete) to the CSS problem. In the proof, the CSS problem is applied to a dense matrix constructed from the three coloring problem. It leaves the complexity of the CSS problem for sparse matrices open. We show that the CSS problem is NP-hard even for sparse matrices.

Theorem 2. The column subset selection problem is NP-hard even if each column has only two nonzero elements.

Proof: By reduction from minimizing NCut to the CSS problem. Suppose the CSS problem is not NP-hard on sparse matrices with two nonzero elements in each column. As shown in (3), the normalized edge matrix has two nonzero elements in each column. Then, this matrix can be used within Algorithm 1 to find the optimal clustering according to NCut in polynomial time. This contradicts the NP-hardness of the NCut problem (Shi and Malik 2000). □

Conclusions

Our main result is showing the equivalence between graph spectral clustering and column subset selection from an edge matrix. To the best of our knowledge, this relationship was not previously known, and it allows mapping results between the two problems. We described some of these results, and we expect many additional results to be discovered.

Acknowledgments

G. Wan would like to thank Melody Wang for her support.

References

Guattery, S.; and Miller, G. L. 1998. On the quality of spectral separators. *SIMAX*, 19: 701–719.

He, B.; Shah, S.; Maung, C.; Arnold, G.; Wan, G.; and Schweitzer, H. 2019. Heuristic search algorithm for dimensionality reduction optimally combining feature selection and feature extraction. *AAAI*, 33: 2280–2287.

Mao, W.; Wan, G.; and Schweitzer, H. 2024. Graph Clustering Methods Derived from Column Subset Selection (Student Abstract). In *AAAI*.

Ng, A.; Jordan, M.; and Weiss, Y. 2001. On spectral clustering: Analysis and an algorithm. *NeurIPS*, 14.

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

Shitov, Y. 2021. Column subset selection is NP-complete. *Linear Algebra and its Applications*, 610: 52–58.

Von Luxburg, U. 2007. A tutorial on spectral clustering. *Statistics and computing*, 17: 395–416.

Wan, G.; and Schweitzer, H. 2021. Heuristic Search for Approximating One Matrix in Terms of Another Matrix. *IJCAI*, 1600–1606.