Optimization of Robust Clustering from Graphs

Fahad Mostafa[1] and Rejuan Haque[2]

1. Texas Tech University, Lubbock, Texas and 2. Ohio State University, Columbus, Ohio.

Introduction

- Lets $G = (V, E)$ be the graph and in spectral clustering, the affinity, and not the absolute location (i.e. k-means), we generate a similar graph (e.g. KNN graph) for all the data points and determines what points fall under which cluster.
- It creates embedded data points in the low dimensional space, in which the cluster is more obvious with the use of the eigenvectors of the graph Laplacian, however in case of most commonly used clustering algorithm such as K-means (Dhillon et al. 2007), is applied to partition the embedding.
- K-means is non convex optimization problem so it has problem with local maxima.
- In spectral clustering (Roxborough & Sen 1997, Shi & Malik2000, Meila & Shi 2001, Ng et al. 2002); Ncut on real dataset, finding eigenvector from big similarity matrix is slower.
- Power iteration clustering (PIC) (Lin and Cohen 2010) on large data sets is 1000 times faster but having defective eigenvalues and still slower.
- We propose Shift Invert with Rayleigh Quotient Clustering (SIRQC) for increasing accuracy of eigenvalues. However when ratio of eigenvalues is nearly one, it gives clustering besides the max eigenvalue of similarity system. It is way more faster.
- **Other methods:** Matrix powering (Zhou & Woodruff 2004), diffusion maps (Lafon & Lee 2006), Gaussian blurring mean-shift (Carreira-Perpinan 2006)

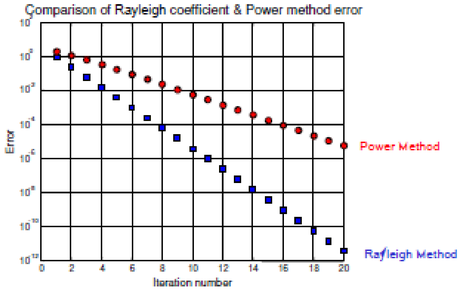


Fig 2- Model Comparison

Method

- Graph G with n vertices. Laplacian matrix is $n \times n$; $L_g = D_g - A_g$,
- $A_g = A_{ij}$ is the Adjacency matrix and D_g is the diagonal matrix of vertex degrees with w_{ij} weights of the edge E
- **MAIN IDEA OF SC**

L_g has eigenvalues $\{\lambda_1, \lambda_2, \dots, \lambda_n\}$ where $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$; eigen vector $\{x, x_2, \dots, x_n\}$; $\lambda_2 > 0$, and λ_2 is the algebraic connectivity M.

- Bi–partitioning of the clusters CL_1 and C_2, and assign nodes with $x_2(i) > 0 \; to \; CL_1$; $x_2(i) < 0 \; to \; CL_2$. cluster : $CL_1 = \{i \,|x_2(i) > 0\}$ and $CL_2 = \{i \,|x_2(i) < 0\}$.

- **Algorithm: Shift Invert with Rayleigh Quotient Clustering (SIRQC)**

Row-normalized affinity matrix W and the number of clusters p and looking for clusters $C_1, C_2 \dots C_p$; Pick $\hat{\vartheta}$ which is nonzero, dynamically choose shifts for shift invert steps using Rayleigh quotients

- Find $\lambda_{k+1} = \frac{\vartheta^{(k)*} W \vartheta^{(k)}}{\vartheta^{(k)*} \vartheta^{(k)}}$

Set $\vartheta^{k+1} = \frac{(W - \lambda_{k+1})^{-1} \vartheta^{(k)}}{\|(W - \lambda_{k+1})^{-1} \vartheta^{(k)}\|}$

Set $\zeta^{t+1} = \left|\vartheta^{k+1} - \vartheta^{k}\right|$;

Stop when $|\zeta^{t} - \zeta^{t-1}| \approx 0$

- Use k-means to cluster points on ϑ^k and get clusters $C_i, i = 1{:}p$
- Convergence: considering only 2 eigenvalues $r(\vartheta^k)$

$$= \lambda_1 \left(1 + \frac{w_2^2}{w_1^2}\left(\frac{\lambda_2}{\lambda_1}\right)^{2n+1}\right)\left(1 + \frac{w_2^2}{w_1^2}\left(\frac{\lambda_2}{\lambda_1}\right)^{2n}\right)^{-1}$$

- Proportional errors, decays with successive iteration as $\left(\frac{\lambda_2}{\lambda_1}\right)^{2n}$ which is quadratic convergence and must faster then PIC.

- Consider m data points in the graph/network with n features i.e. $X \epsilon R^{m \times n}$
- Minimize $\varphi = \frac{1}{m}\sum_{i=1}^{m} \|x^i - \mu_{C^i}\|$;

x^i row of X, μ_{C^i} cluster centroid of cluster c to which x^i been assigned.

Numerical Results

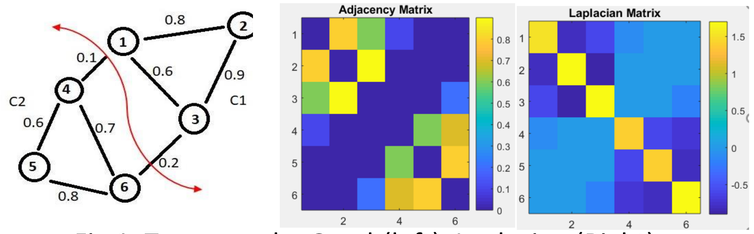
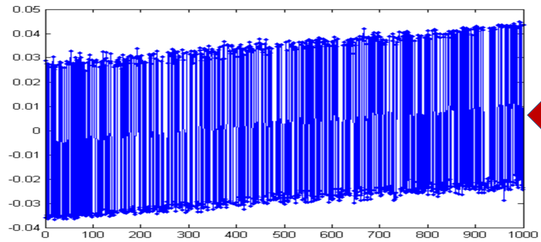


Fig 1. Toy example, Graph(left), Laplacian (Right)



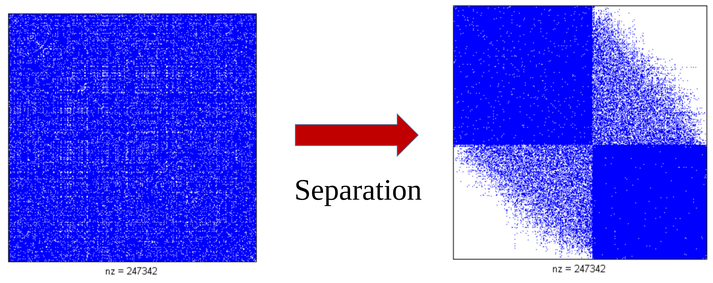


Fig 3. Simulated data, Random Permutation of all vertices

- SIRQC is very handful for very large dataset with sparse setting, such as block-stochastic networks. We use simulated data here to see clusters.

Runtime Comparison Table using Simulated Data (Time in millisecond)

Nodes	Edges	NCuts	PIC	SIRQC
1000	10000	2050	5	1
5000	250000	149789	12	3
500000	500000000	ROM	14900	1460

- ROM -> Run out memory

- **Application(several):** Our Focus on Single cell analysis in Biology. Apply in Graph-based community detection to identify cell types. Abstractly, our graph would just be composed of all our cells as vertices. Data is simulated.

Conclusion

- SIRQC– a variant on the power method faster than PIC because of quadratic convergence.
- different approach from the spectral methods mentioned above.
- Very rapid convergence is guaranteed and no more than a few iterations are needed in practice to obtain a reasonable approximation. The Rayleigh quotient iteration algorithm converges cubically for Hermitian or symmetric matrices, given an initial vector that is sufficiently close to an eigenvector of the matrix that is being analyzed.
- Future Work: Apply on more real data sets for experimental results.
- Testing Scalability just like PIC, also its works very well on eigenvalue weighting .

Key References

- Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. On spectral clustering: Analysis and an algorithm. Advances in neural information processing systems, 2:849-856, (2002)
- Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. IEEE Transactions on pattern analysis and machine intelligence, 22(8):888-905, (2000)
- Lloyd N. Trefethen and David Bau, III, Numerical Linear Algebra, Society for Industrial and Applied Mathematics, (1997)
- Lin and Cohen , Power iteration clustering, (2010)

Fig 4- Five very clearly distinct cell types