Optimization of Robust Clustering from Graphs

Fahad Mostafa¹ and Rejuan Haque²

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

Introduction

- Let $G = (V, E)$ 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 (Roxburgh & Sen 1997, Shi & Malik 2000, Meila & Shi 2001, Ng et al. 2002); Neat 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 (Laflon & Lee 2006), Gaussian blurring mean-shift (Carreira-Perpinan 2006)

Method

- Graph $G$ with $n$ vertices. Laplacian matrix is $n \times n$: $L_G = D_G - A_G$.
- $L_G$ 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 $\{\lambda_1, \lambda_2, ..., \lambda_n\}$ where $0 = \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$, eigen vector $\{x_1, x_2, ..., x_n\}$ $\lambda_i > 0$, and $\lambda$ is the algebraic connectivity $M$.
- Bi-partitioning of the clusters $C_1$ and $C_2$, and assign nodes with $x_i(\ell) > 0 \rightarrow C_1$; $x_i(\ell) < 0$ to $C_2$. cluster $C_1 = \{i | x_i(\ell) > 0\}$ and $C_2 = \{i | x_i(\ell) < 0\}$.
- Algorithm: Shift Invert with Rayleigh Quotient Clustering (SIRQC)

Row-normalized affinity matrix $W$ and the number of clusters $p$ and looking for clusters $C_1, C_2, \ldots C_p$:
- Pick $\theta$ which is nonzero, dynamically choose shifts for shift invert steps using Rayleigh quotients
- Find $\lambda_{k+1} = \frac{\delta^{3} w \delta}{\delta^{3} \delta^{2}}$.
- Set $\theta_{k+1} = \frac{\lambda_{k+1} - \lambda_{k}}{\|W^{1/2} \lambda_{k}^{1/2} \|^2}$.
- Set $\zeta_{k+1} = \frac{\theta_{k+1} - \theta_{k}}{\|W^{1/2} \lambda_{k}^{1/2} \|^2}$.
- Stop when $\|\zeta - \zeta_{k-1}\| = 0$.
- Use k-means to cluster points on $\theta_k$ and get clusters $C_i, i = 1:p$
- Convergence: considering only 2 eigenvalues $r(\theta_k) = \lambda_1 \left(1 + \frac{w_{2}^2}{w_{1}^2} \lambda_2^2 \right)^{-1}$
- 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. $X \in \mathbb{R}^{m \times n}$
- Minimize $\varphi = \frac{1}{2} \sum_{c=1}^{C} \|x_c - \mu_c\|^2$.
- $x_c$ row of $X$, $\mu_c$ cluster centroid of cluster $c$ to which $x_c$ has been assigned.

Numerical Results

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

Second smallest eigenvector

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)

<table>
<thead>
<tr>
<th>Nodes</th>
<th>Edges</th>
<th>NCuts</th>
<th>PIC</th>
<th>SIRQC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>10000</td>
<td>2090</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>5000</td>
<td>25000</td>
<td>149789</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>500000</td>
<td>500000000</td>
<td>ROM</td>
<td>14900</td>
<td>1460</td>
</tr>
</tbody>
</table>

- ROM -> Run out memory

Key References


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.

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.