Spectral Clustering of Chinese Herbal Medicine Network

Guangyu Tong[1], Liqi Feng[2] (NetID: gt57, lf109)

[1]Department of Sociology [2]Department of Biostatistics & Bioinformatics

Motivation

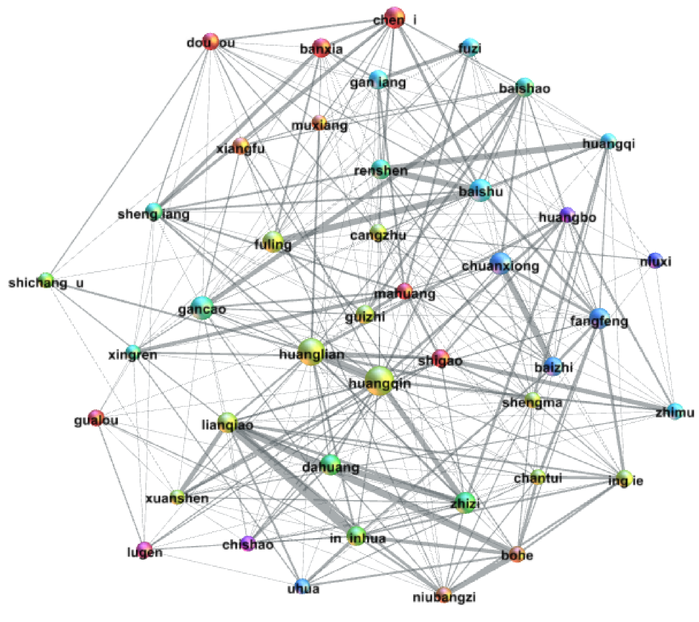
Traditional Chinese Herbal Medicine (TCHM) is effective in treating many diseases. Yet the nature of TCHM is not well-understood. A prescription of TCHM often includes a compound mixture of 3-10 herbal ingredients and its efficacy highly depends on the synergy effect of herbal ingredients. The challenges in understanding the synergy effect of herbal ingredients arouse because of the complicated interrelationship structure.

Spectral Clustering is widely used in problem where the data is arranged in a complex and unknown shape (Zelnik-Manor and Perona, 2005). In this project, the interrelationship between different herbal ingredients is investigated by Spectral Clustering.

- The interrelationship between herbal ingredients is represented by a communication network where ingredients shown on the same prescription are connected;
- Normalized symmetric graph Laplacian is used to investigate the communication network.

TCHM prescription data

Figure 1: Medicine network of 603 prescriptions ($V = 213$, $E = 2633$). Only degree > 25 is presented.



Symmetric Graph Laplacian Spectral Clustering

- Weighted *adjacency matrix* of a communication network $W = (w_{ij})_{i,j=1,...,n}$ where ingredients ν_i and ν_j disconnected $\rightarrow w_{ij} = 0$. Otherwise, w_{ij} is the number of connections ingredients ν_i and ν_j have based on the specified number of nearest neighbors
- graph Laplacian $L = D - W$ where D is a diagonal degree matrix, $D_i = \Sigma_{j=1}^{n} w_{ij}$
- Herbal ingredients have similar degrees of edges as a prescription often contains a lot of ingredients
- Symmetric graph Laplacian $L_n = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$

Number of clusters is determined by the eigengap heuristic

- The number of small eigenvalues (≈ 0) of L_{sym} indicates the number of clusters in the original communication network
- Determine the appropriate number of clusters by finding the eigenvalue λ_k of of L_{sym} and λ_{k+1} is relatively large. k is the number of clusters used in the analysis

Algorithm (Ng, Jordan and Weiss, 2002)

1) Construct a similarity graph by connecting two herbal ingredients based on the distance in the communication network

2) Compute the normalized Laplacian L_{sym}.

3) Compute the first k eigenvectors (ascending order) $u_1, ..., u_k$ of L_{sym}.

4) Let $U \in \mathbb{R}^{n \times k}$ contains $u_1, ..., u_k$ as columns.

5) Form the matrix $T \in \mathbb{R}^{n \times k}$ from U by normalizing the rows to norm 1 ($t_{ij} = u_{ij} / (\Sigma_k u_{ik}^2)^{\frac{1}{2}}$).

6) For $i = 1, ..., n$, let $y_i \in \mathbb{R}^k$ be the vector corresponding to the i^{th} row of T.

7) Clusters the points $(y_i)_{i=1,...,n}$ with the k-means algorithm into clusters $C_1, ..., C_k$.

The above analysis was done using R (version 3.2.2). Package igraph is employed to construct the communication network and package kknn is used for spectral clustering. Package igraph is used to generate method comparison table.

Results

Figure 2: Eigenvalue plot of the Laplacian matrix for adjacency matrix connecting the nearest 4 ingredients. Two arrows indicate the optimal number of cluster for the algorithm (eigengap heuristic)

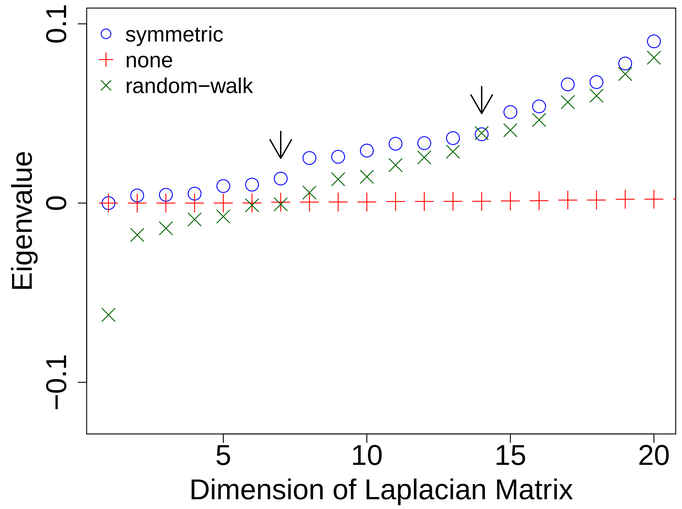


Table 1: Comparison of performance between different clustering methods.

Method	# Clusters	Misclassification Rates (%)
Lap-Symm	14 ($K = 4$)	(14.3, 34.2)
Lap-Symm	4 ($K = 10$)	(38.8, 10.0)
Lap-None	12 ($K = 3$)	(40.0, 36.8)
Betweenness	84	(8.0, 23.5)
InfoMap	68	(0.0, 50.0)
Label Propogation	1	(47.5, 100.0)
Lead Eigenvalue	10	(13.0, 30.6)
Multilevel	6	(13.0, 30.6)
Stat Mechanics	7	(13.0, 30.6)
Short Ran-Walk	14	(17.9, 25.8)

misclassification rates is computed as the type I error and type II error

Conclusion

1) Graph Laplacian spectral clustering achieves reliable performance (small type I and type II errors).

2) Graph Laplacian spectral clustering outperforms other methods in probing wrong prescription combination.

3) Symmetric approach outperforms the unnomalized approach and the random walk approach.

4) The performance of clustering can be improve with a more complete data set with disease information.