Network Theory Methods and Measures

Violet Liu

1 Introduction

In what follows, we will discuss the methods and measures we planned to use to analyze **PPI networks** and solve biological problems given to us by the BCMB team.

2 Louvain algorithm

2.1 Background

The Louvain algorithm is a kind of algorithm, which optimizes **Modularity** on multiple levels. Modularity is generally used to measure the quality of the results of the community discovery algorithm, showing the strength of particular communities. The Louvain algorithm uses iterative calculations to improve the modularity in the community. If a node is added to the community within one of its neighbors, and the modularity of that community is optimized, the node is accepted into that community.

2.2 Formulation

Let the communitiesâ modularity be denoted by Q. If Q becomes constant, the iteration is stopped. The formula for Q is as follows:

$$Q = \frac{1}{2m} \sum_{i,j} B_{ij} S(C_i, C_j)$$

$$B_{ij} = A_{ij} - \frac{1}{2m} k_i k_j$$

Above, A_{ij} is the sum of node i and node j, and all weights of edge equals 1 when it is a unweight graph. Meanwhile, $A_{ij} - \frac{1}{2m}k_ik_j$ shows the sum of weights of all edges which is connected with node $i.c_1$ and c_2 is the corresponding community of node i and node j, $\frac{1}{2m}k_ik_j$ means the number of edges.

Now, if we put node i in the community which has its neighbor – node j, the formulation will become:

$$Q = \sum_c e_c - a_c^2$$

Therefore, Q can be also understood as the weight of inner edge of community minus the sum of the weights of the edges which connected all nodes in the community. In an undirected graph it is better to explanate, that is, the degree of the internal edges minus the sum of degrees of the nodes in the community.

2.3 Algorithm

The Louvain algorithm consists of two stages. Primarily, it will visit all nodes in the network, try to add a single node to the community, and find a community which can be maximized the modularity by this node. This step will be stopped until all nodes no longer change. The results of the stage one will be solved in step two. In this process, every small community will be regarded as a new node, then, a new network structure is produced. meanwhile, in this node transformed from small community, the current weight of edge is the sum of the edge weights of all original nodes. Finally, it will repeat these two steps until Q is stable.

2.4 Conclusion

To sum up, aim of Louvain algorithm is maximizing modularity Q. This model can support strategy to improve, to continuously construct a community with internal clustering but external connections sparsing.

3 Spectral Clustering

3.1 Background

Spectral clustering including undirected graphs, linear algebra and matrix analysis, is an algorithm evolved from graph theory, which has been widely used in clustering, currently. It works by treating all data as points in space, then, edges can connect these points. If two points are close, the weight of edge between these two points is quite high. In contrast, the edge weight between two points will be lower, when they are farther apart than others. Furthermore, the clustering can be achieved, by cutting the graph. Since the sum of the edge weights between the subgraphs will be as low as possible, with a possibly highest sum of weight of the inner edges.

3.2 Formulation

Generally, speaking, the main attention points of spectral clustering are the generation method of similarity matrix ($s_{ij} = ||x_i - x_j||_2^2$), Graph Laplacians($f^T L f = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2$), the method of cutting graph ($cut(A_1, A_2, ...A_k = \frac{1}{2} \sum_{i=1}^{k} W(A_i, \bar{A}_i))$), and the final clustering method (*k-means clustering*).

3.3 Algorithm

There is a simplify process of Spectral Clustering:

- Input: a sample set $D = (x_1, x_2, ..., x_n)$, the method to generate similarity matrix , dimension k_1 after dimensionality reduction, clustering method, dimension k_2 after clustering;

1) Construct the sample similarity matrix S;

2) Build adjacency matrix W and degree matrix D;

3) Calculate the Laplacian matrix L, geting a a standardized Laplacian matrix $D^{-1/2}LD^{-1/2}$;

4) Find the eigenvector f corresponding to the minimum of eigenvalues k_1;

5) Formed a $n \times k_1$ dimensional eigenmatrix F;

6) Create n samples in F, clustering using the method in *Input*, the clustering dimension is k_2;

- Output: a cluster partition $C(c_1, c_2, ...c_{k_2})$.

3.4 Conclusion

Spectral clustering is based on the data similarity matrix. It defines the optimized objective function of subgraph, and makes improvements (method of cutting graph), introduces indicator variables, and converts the problem into finding an optimal matrix.

4 The Markov Cluster Algorithm (MCL)

4.1 Background

MCL is a fast and scalable clustering algorithm based on graph. Supposing there is a certain area A, which is dense, in a sparse graph G. Now, if we randomly walk k steps in A, with quite large probability in A, it means that A has many k-length paths. On the other words, if it can walk randomly k steps in the graph and the probability in this area is very high, the area can be said as a cluster. However, the next step of random walk is only related to the current node, which is called a Markov random walk process. In this process, **Expansion** and **Inflation** will continually alternate. Finally, the phenomenon of clustering will gradually appear, until clustering finished.

4.2 Definition

- Expansion can strengthen the connection between different areas;
- Inflation continuously differentiates the connection between points.

4.3 Algorithm

- Input: a non-fully connected graph, the parameter e during **Expansion** and the parameter r during **Inflation**;

1) Construct the adjacency matrix W;

2) Add a self-loop;

3) Normalized the probability matrix;

4) **Expansion** working, the matrix is multiplied e power each time;

5) **Inflation** each time the elements in the matrix are multiplied r power, and then standardized it;

6) Repeat *4)* and *6)* until it is stable;

- Output: convert the result matrix to clusters.

4.4 Conclusion

The parameter of Inflation will affect the clustering. Generally, when r increases, its particle size will decrease. In addition, the value of e will also affect the number of clusters, which is more obvious in the large-diameter graph. Because points in remote areas has a less connected to center of cluster than the points which is close with the center. It seems that points will automatically change to other cluster or lead to internal differentiation problems in cluster.

5 Other Measures

There are list the main measures we planed to used, they are:

- **Degree Centrality:** in an non-directed network, the degree of a node can be used to measure centrality, because the important nodes are connected to more nodes others. In other words, when a node has stronger influence, this means that it is related to more nodes at the same time. Thus, we can find the nodes which can affect more other nodes in a PPI network.

- **Betweenness Centrality:** in the network, when two non-adjacent nodes interact, they must depend on other nodes, especially those nodes which are on the path between those two nodes. At the same time, the nodes on their path can control and restrict the interaction between them. Therefore, if a node is located on the multiple shortest path between other nodes, this node is the core node with a large intermediary centrality. When many members rely on this core node to contacts or low-cost contacts with others, in a network , it can control and restriction on other members. It will have a great impact on the transfer of the whole graph.

- **Eigenvector Centrality:** the importance of a node not only depends on the number of nodes connected with it (the degree of the node), but also depends on the importance of its neighbor nodes. It means that, the more important the nodes connected to a same node, this node will become more important. Let x_i be the importance metric value of node v_i, $EC(i) = x_i = c \sum\limits_{j=1}^{n} a_{ij} x_j$. In addition, x is the eigenvector corresponding to the eigenvalue c_1 of matrix A, $Ax = \lambda x$. The centrality of the i^{th} node is equal to the i^{th} element in the feature vector.

- **Katz Centrality:**the principle of Katz Centrality is similar of eigenvector centrality, however, each node needs to be assigned an initial centrality value, β, and the top 10 nodes with the largest value can be discharged through the results. There is its formulation: $x = \beta(I - \alpha A)^{-1} \cdot 1$.

- **PageRank** is a variant of eigenvector centrality, meanwhile, it is also an index to measure the importance of nodes in a directed network. Because in the network, there is a direction when one node moves from to another. In other words, node A can jump to the node B through the link, but it does not mean that node B can return back to node A. Convert these directions into edges to form a directed graph.Therefore, the degree of influence of a node is equal to the sum of the weighted influence of all nodes in the chain set, and the formula is expressed as: $PR(u) = \sum_{u \in B_i} \frac{PR(u)}{L(u)}$.

References

[1] Chatterjee M. âIntroduction to Spectral Clusteringâ. Great learning, Aug 16, 2020, https://www.mygreatlearning.com/blog/introduction-to-spectral-clustering/

[2] Disney A. âSocial network analysis 101: centrality measures explainedâ. Cambridge Intelligence, Jan 2, 2020, https://cambridge-intelligence.com/keylines-faqs-social-network-analysis/

[3] Golbeck J. âNetwork Structure and Measuresâ. ScienceDirect, 2013, https://www.sciencedirect.com/topics/computer-science/centrality-measure

[4] Micans. âMCL - a cluster algorithm for graphsâ. Micans.org, 2020, https://micans.org/mcl/

[5] Neo4j. â6.1. The Louvain algorithmâ. Graph Data Science Library (GDS), 2020, https://neo4j.com/docs/graph-algorithms/current/algorithms/louvain/

[6] Rita L. âLouvain Algorithmâ. Towards data science, Apr 10, 2020, https://towardsdatascience.com/louvain-algorithm-93fde589f58c

[7] Shaw A. âUNDERSTANDING THE CONCEPTS OF EIGENVECTOR CENTRALITY AND PAGERANKâ. Strategic planet, Jul 13, 2019, https://www.strategic-planet.com/2019/07/understanding-the-concepts-of-eigenvector-centrality-and-pagerank/