ISSN 2616-5775 Vol. 5, Issue 12: 82-88, DOI: 10.25236/AJCIS.2022.051212

An Improved Algorithm for Label Propagation Based on Rough Core

Laizong Huang[1,*], Fei Liu[1]

[1]School of Software, Jiangxi Normal University, Nanchang, 330022, China
**Corresponding author: 13065119131@163.com*

***Abstract:** Complex networks are a hot topic in social science research today. Analyzing the structure of the relationships embedded in them through community discovery techniques is an important tool for the study of complex networks. To address the randomness problem of label propagation in label propagation algorithms, this paper proposes an improved algorithm for label propagation based on rough kernels. The algorithm first performs the extraction of rough kernels and assigns unique labels to each rough kernel to achieve fast label preprocessing; then intervenes in the label propagation process through the ranking of node influence to reduce its randomness and make the division results more stable; finally, the community division is performed by the distribution of network labels. Experiments are conducted in real networks, and the results show that the combined performance of the method in this paper is better than the traditional labeling algorithm in terms of modularity and NMI metrics, and the proposed improved algorithm leads to improved community division results.*

***Keywords:** complex networks, community discovery, label propagation, node influence*

1. Introduction

Complex networks are a hot topic of social science research today, and many complex systems can be represented as networks, such as citation networks, interpersonal networks, mobile Internet, etc. Since these networks are rich in data information, analyzing the structure of the relationships embedded in them through community discovery techniques is an important tool for the study of complex networks. For example, community discovery can be used to mine the economic activities of a company with other industries, so that the economic dynamics of this company can be accurately discovered. At this stage, complex networks show huge and dynamic trends, and community discovery can provide important reference information for urban planning, Internet development, public opinion analysis, and hazard protection.

Complex networks represent a collection of complex systems, which are composed of individuals through intricate connections, which contain rich structural information and attribute properties, and they are characterized by small world [1], scale-free [2], and community structure [3], which are the basis for community structure discovery. Community discovery covers technical knowledge in many fields such as graph theory, statistics, and data mining, which can be used to detect community structures in complex networks and thus help people understand the composition, development, and evolution of various relationships in the real world. Usually, community discovery views things with connections in complex networks as nodes and such connections as directed or undirected edges, and then constructs real-world network models in the form of topological diagrams, so that relevant community discovery techniques can be used to analyze, derive and implement this information. Community discovery is an important method for studying the structural characteristics of complex networks and an important tool for analyzing human connections, which can accurately and efficiently uncover deeper interpersonal relationships.

Communities have the distinct characteristic of being strongly linked between nodes within the community and sparsely linked between nodes outside the community. The strength of such ties is fundamental to what constitutes a community, so the study of community discovery requires an analysis of the linkage relationships involved.

The classical hierarchical clustering algorithm Girvan and Newman proposed the GN algorithm [4], which initially treats the community as a whole and then achieves the division of the community by removing the edge with the largest number of edge meshes. Such methods are also called disaggregated

ISSN 2616-5775 Vol. 5, Issue 12: 82-88, DOI: 10.25236/AJCIS.2022.051212

hierarchical clustering, but because such strategies lack the evaluation of discovery results, hierarchical clustering based on local extensions is more popular among researchers. The local extension method is the reverse of split hierarchical clustering, i.e., each node is first considered as a community, and then the merging of communities is performed by extension, and finally the division of communities is achieved. For example, the CDHC algorithm proposed by Yin et al [5], which first finds the global central node, then calculates the link strength between nodes, and finally merges small communities into large communities based on that link strength. Similarly, the literature [6] defines the global central node as a leader node, and merges the leader node with the largest similar node by calculating the structural similarity among members for iterative update to detect the implied community structure. Although the central or leader node is beneficial to improve the accuracy of community discovery, the cost and arbitrariness of such inexpensive links lead to the initial search for the central or leader node is not easy to achieve due to the existence of certain dependencies when nodes propagate or exchange information in complex networks where each node has rich interaction information. Therefore, Jiang et al [7] focused on the influence of other neighboring nodes and judged the formation of the initial community by calculating the local influence, thus proposing a framework for maximizing the influence calculation and improving the efficiency of the discovery of the maximally influential nodes.

Graph partitioning algorithms first partition the nodes in a social network into many unrelated groups, and then discover communities by finding the partitioned, least edge-connected groups. One of the most common graph segmentation methods is the spectral analysis method [8], the core of the spectral analysis method is to construct the Laplace adjacency matrix and then calculate the eigenvectors of the matrix so as to perform object segmentation or pattern analysis based on the positive and negative cases in the eigenvectors, so the spectral analysis method can be applied to social networks to obtain more ideal segmentation results. However, since using the positive and negative cases as reference criteria will reduce the efficiency of community division, Capocci et al [9] proposed a community discovery method based on semantic similarity for spectral leveling, which uses the semantic properties of social networks as community division criteria to perform community analysis more effectively. Shen et al [10] addressed the problem of quadratic binary division in spectral leveling and proposed a distance-based community discovery method for spectrum analysis, which groups a set of objects using the spectrum of paired distance matrices, thus associating points in the metric space with defined distances and further improving the community partitioning results.

The modularity degree was first proposed by Newman[11], which indicates the quality of a community's division, and if the modularity degree is larger, it indicates the higher quality of community division and the more it indicates a distinct community structure. Module degree is an important evaluation index for community discovery, and many scholars have used module degree to discover communities, such as LM algorithm [12], simulated annealing algorithm [13], and CONCLUDE algorithm [14], which are used to find the maximum module degree in the process of community division, so as to subdivide the community and thus obtain the community structure under the optimal module degree. Although these methods have good reference value, this division has serious resolution limitation problem [15], so many scholars target improved modularity for community detection, such as Mairisha et al [16] proposed an improved MC modularity to measure the modularity of a particular community by the grinding coefficient to get a more accurate community division. Chen et al [17] proposed an adaptive modularity calculation method that can combine the advantages of different modularity, and revealed the effectiveness and superiority of multi-resolution modularity in community detection by applying modularity to various synthetic networks and real-world networks. Qiao et al [18] proposed an overlapping community discovery algorithm in large data of complex networks, which is based on modularity for clustering, designed a new way of updating modularity, and then indexed the modularity increments by balanced binary trees to get a better community delineation. To perform larger scale community discovery, Qiao et al [19] performed parallel community discovery via Hadoop and used the modularity increment to effectively solve the problem that traditional community discovery algorithms cannot handle large scale datasets.

Label propagation is a strategy in the field of machine learning. The LPA algorithm proposed by Zhu et al. in 2002 [20], which is a graph-based semi-supervised learning method, has the basic idea of using the label information of labeled nodes to predict the label information of unlabeled nodes. the RAK algorithm proposed by Raghavan et al. in 2007 [21] applied label propagation to community detection for the first time. The advantage of LPA is its simple procedure and linear time complexity. Two core steps of LPA are to assign different labels to the nodes in the network and to select the labels of the maximum number of nodes around that node for propagation in a random iteration process. The label propagation algorithm is a heuristic strategy-based algorithm that does not rely on prior knowledge and does not require setting an objective function, which has achieved better results in many real networks. However, the label propagation process of the traditional label propagation algorithm considers that each neighbor node of the node being updated has an equal impact on that node, and the connection relationship between the neighbor node and that node is not considered, which can easily lead to arbitrary propagation of labels among different communities, which in turn affects the accuracy of the label propagation algorithm.

For the problem of label propagation algorithm label propagation randomness, an improved rough kernel-based label propagation algorithm CLPA is proposed in this paper and verified by experiments.

2. Label propagation algorithm

2.1. Description of the label propagation algorithm

The main idea of the label propagation algorithm is that the label of a node depends on the label values of all its neighbor nodes and the nodes with the same label belong to the same community. The label propagation algorithm is described as follows.

(1) Initialization. Assigning each node a label that uniquely represents the community to which it belongs;

(2) Number of iterations t = 1;

(3) Random ordering of all nodes to generate a sequence of nodes X;

(4) Label update. For each node in the node sequence, update the label of that node with the one that appears most frequently in the labels of its neighboring nodes, and if there is more than one label with the highest frequency in the labels of the neighboring nodes, choose a label from these labels at random as the label of that vertex;

(5) If the labels of all nodes no longer change, the algorithm stops; otherwise, t = t + 1 and returns to step 4;

(6) All vertices with the same label are grouped into one community.

2.2. Problems of label propagation algorithm

The LPA algorithm seems to be perfect, but the LPA algorithm treats all neighboring nodes equally in label propagation, and this method of randomly selecting a neighboring label may lead to the random propagation of labels among different communities. The update order of nodes is random, which leads to the diversity of final results and naturally decreases the stability; during the update process, the difference in importance between nodes is ignored, and the nodes with small influence may in turn influence the nodes with larger influence, leading to a decrease in the accuracy of the results.

3. Rough core-based label propagation algorithm

3.1. Community Discovery Related Concepts

The network can be represented by an undirected graph G = (V, E), where V denotes the set of nodes and E denotes the set of edges. The basic concepts used in this paper are described below.

Definition 1 (Community Neighbor Set) The community neighbor set $N_s(C)$ denotes the set of nodes that have directly connected edges to the community C.

$$N_s(C) = \cup_{v \in C} \tau(v) - C \tag{1}$$

$$\tau(v) = \{u : u \in V, (v,u) \in E\} \tag{2}$$

Where C in equation (1) represents a community and equation (2) is the definition of the set $\tau(v)$ of neighbors of node v.

Definition 2 (Influence score) The influence score $I_{score}(v)$ of node v is defined as:

$$I_{score}(v) = k_v \times \sum_{u \in \tau(v)} (k_u \times J_{uv}) \tag{3}$$

Where, k_v is the degree of node v. A larger $I_{score}(v)$ represents the greater influence that node v

ISSN 2616-5775 Vol. 5, Issue 12: 82-88, DOI: 10.25236/AJCIS.2022.051212

has in the network.

3.2. Label propagation algorithm improvement ideas

As analyzed earlier, the random propagation of labels among community edges is an important reason affecting the accuracy of the LPA algorithm, then the accuracy and stability of community segmentation can be improved if the random propagation of edge node labels among different communities can be suppressed when the labels of community edge nodes are updated.

To this end, the concept of influence score can be introduced in the label updating process of node x. If the neighbor node y of x has a larger influence score representing the node, but node x is more closely associated with node y. When there are multiple labels with the highest frequency among the neighboring node labels of a node in label propagation, the label of the node with the largest node influence score is selected from these labels as that vertex label.

Meanwhile, in order to improve the accuracy of the LPA algorithm, the algorithm in this paper preprocesses the initial network and proposes a rough kernel extraction method, the specific process of which is as follows.

(1) Initially assigning a unique label to each node;

(2) Calculate the degree of all nodes and sort all nodes from smallest to largest;

(3) Iterate through all nodes. For each node in the sequence, in order to find a new rough kernel, first find a "free" vertex k with maximum degree, and then find a second "free" node v with maximum degree among the neighboring nodes of k. Here the own node means a node that does not belong to any rough kernel yet. Based on the two selected nodes, the vertex with the smallest degree is iteratively added from the set of common neighbor nodes in the current rough kernel to find the smallest and largest cluster as the rough kernel;

(4) If the number of rough kernels no longer changes, stop the preprocessing; otherwise, t=t+1, and return to the third step.

3.3. Description of the algorithm

According to the previous analysis, the process of the CLPA algorithm is as follows:

(1) Rough kernel extraction;

(2) Initialization. Assigning a label to each rough kernel that uniquely represents the community to which it belongs;

(3) Number of iterations t = 1;

(4) Calculate the influence scores of all nodes, sort them and generate a sequence of nodes X;

(5) Label update. For each node in the node sequence, update the label of that node with the label of the node with the highest frequency in the labels of its neighboring nodes, and if there are multiple labels with the highest frequency in the labels of the neighboring nodes, select the label of the node with the highest influence score of the neighboring nodes from these labels as the label of that vertex;

(6) If the labels of all nodes no longer change, the algorithm stops; otherwise, t = t + 1 and returns to step 4;

(7) All vertices with the same label are grouped into one community.

4. Experiment and analysis

To validate the CLPA algorithm, two real networks commonly used for community discovery algorithm evaluation are selected as test datasets in this paper, one is the Zachary Karate Club network with relatively simple network and fewer nodes, and the other is the American University American Football Tournament network with more nodes and complex network structure. The different algorithms were also tested several times with the same dataset, considering the stability problem conducted by the LPA algorithm.

The comparison data of 20 consecutive experiments were randomly selected to demonstrate the effect

ISSN 2616-5775 Vol. 5, Issue 12: 82-88, DOI: 10.25236/AJCIS.2022.051212

of changes brought by the CLPA algorithm, and the experimental data are shown in Tables 1 and 2.

Table 1: Experimental data of Karate data set

Experiment Number of times	Number of communities		Q-value		NMI-value	
	CLPA	LPA	CLPA	LPA	CLPA	LPA
1	3	3	0.3653	0.3444	0.7210	0.6085
2	2	4	0.3718	0.3548	0.6772	0.6178
3	2	3	0.3037	0.1328	0.3170	0.2065
4	4	2	0.4033	0.4033	0.5056	0.4420
5	3	2	0.4020	0.3600	0.5684	0.4424
6	3	2	0.3691	0.3715	0.8023	0.6372
7	4	3	0.3993	0.3548	0.6009	0.6178
8	3	4	0.4020	0.3742	0.5684	0.4006
9	4	3	0.3993	0.4020	0.6009	0.5684
10	2	2	0.3718	0.3549	0.6772	0.6372
11	4	4	0.3752	0.3638	0.4513	0.2817
12	2	3	0.3718	0.3544	0.6772	0.6085
13	4	3	0.3743	0.3144	0.6648	0.4085
14	3	3	0.3875	0.3691	0.7053	0.5023
15	4	3	0.4151	0.3151	0.7000	0.600
16	2	1	0.3923	0.3542	0.7324	0.6241
17	3	3	0.3764	0.3991	0.5871	0.6912
18	3	3	0.3947	0.3644	0.6313	0.6163
19	3	4	0.3991	0.3742	0.6912	0.4006
20	4	3	0.4033	0.3.812	0.5656	0.5084

Table 2: Experimental data for the football data set

Experiment Number of times	Number of communities		Q-value		NMI-value	
	CLPA	LPA	CLPA	LPA	CLPA	LPA
1	9	10	0.5571	0.5689	0.8978	0.8079
2	10	11	0.6568	0.5034	0.9019	0.8846
3	10	10	0.6568	0.5437	0.9019	0.8655
4	10	11	0.5571	0.5301	0.8978	0.8269
5	9	11	0.5572	0.5309	0.8981	0.8269
6	10	9	0.5568	0.5117	0.9019	0.7040
7	10	12	0.5571	0.5473	0.8978	0.8648
8	9	10	0.5565	0.5494	0.8971	0.8588
9	10	12	0.5563	0.5483	0.9019	0.8190
10	9	11	0.5571	0.5492	0.8978	0.8339
11	9	10	0.5568	0.5372	0.8971	0.7688
12	9	10	0.5568	0.5326	0.9019	0.8826
13	9	11	0.5606	0.5574	0.8984	0.8971
14	9	10	0.5946	0.5029	0.8946	0.8882
15	9	9	0.5568	0.5491	0.9019	0.8985
16	10	10	0.5571	0.5471	0.8978	0.8580
17	9	10	0.5572	0.5402	0.8981	0.8647
18	10	6	0.5568	0.3346	0.8971	0.6863
19	10	12	0.5568	0.5043	0.9019	0.8854
20	9	9	0.5572	0.5477	0.8981	0.7926

It can be seen that the quality of community division is not linearly related to the number of communities, because even if the number of communities divided before and after is the same, the distribution of nodes within the community is not exactly the same. The randomness of label propagation determines that there is only a very small probability that the exact same division result will occur twice, and this probability becomes lower and lower as the size of the network dataset continues to grow.

Table 1 and Table 2 show the data comparison between CLPA algorithm and LPA algorithm on karate dataset and football dataset respectively. Through the data in Table 1 and Table 2, it can be found that CLPA algorithm is larger than LPA algorithm in terms of module degree and NMI index, which proves

ISSN 2616-5775 Vol. 5, Issue 12: 82-88, DOI: 10.25236/AJCIS.2022.051212

that CLPA algorithm divides better results than LPA algorithm. Through 20 consecutive experiments, it can be found that the interval of module degree size of CLPA algorithm is smaller than the interval of module degree size of LPA algorithm, which proves that CLPA algorithm is more stable compared with LPA algorithm.

5. Conclusion

In this paper, starting from improving the stability of label propagation in label propagation algorithm, we design a label propagation algorithm based on rough kernel by improving the label propagation algorithm, and conduct experiments with the improved algorithm and the traditional label propagation algorithm in karate club network and American college American soccer league network, respectively. The experimental results show that the community division results of the improved algorithm proposed in this paper are closer to the experimental situation. the CLPA algorithm not only improves the stability of community division, but also can improve the accuracy of community division. Meanwhile, in order to adapt to the division of overlapping communities, an overlapping community discovery strategy will be introduced to achieve the mining of overlapping communities and further enhance the practicality of the algorithm.

References

[1] Barabasi A L, Albert R. Emergence of scaling in random networks [J]. Science, 1999, 286 (5439):509-512.
[2] M. Girvan and M. E. J. Newman, ‘‘Community structure in social and biological networks,’’ Proc. Nat. Acad. Sci. USA, vol. 99, no. 12, pp. 7821–7826, Jun. 2002.
[3] Raghavan Usha Nandini and Albert Réka and Kumara Soundar. NIAr linIAr time algorithm to detect community structures in large-scale networks. [J]. Physical review. E, Statistical, nonlinIAr, and soft matter physics, 2007, 76(3 Pt 2): 036106.
[4] Girvan M, Newman M E J. Community structure in social and biological networks [J]. Proceedings of the national academy of sciences, 2002, 99(12): 7821-7826.
[5] Yin C, Zhu S, Chen H, et al. A method for community detection of complex networks based on hierarchical clustering [J]. International Journal of Distributed Sensor Networks, 2015, 11(6): 849140.
[6] Helal N A, Ismail R M, Badr N L, et al. Leader-based community detection algorithm for social networks [J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2017, 7(6): e1213.
[7] Jiang F, Jin S, Wu Y, et al. A uniform framework for community detection via influence maximization in social networks[C]//2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014). IEEE, 2014: 27-32.
[8] Pothen A, Simon H D, Liou K P. Partitioning sparse matrices with eigenvectors of graphs [J]. SIAM journal on matrix analysis and applications, 1990, 11(3): 430-452.
[9] Capocci A, Baldassarri A, Servedio V D P, et al. Friendship, collaboration and semantics in Flickr: from social interaction to semantic similarity[C]//Proceedings of the International Workshop on Modeling Social Media. 2010: 1-4.
[10] Shen G, Ye D. A distance-based spectral clustering approach with applications to network community detection [J]. Journal of Industrial Information Integration, 2017, 6: 22-32.
[11] Newman M E J. Fast algorithm for detecting community structure in networks [J]. Physical review E, 2004, 69(6): 066133.
[12] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks [J]. Journal of statistical mechanics: theory and experiment, 2008, 2008(10): P10008.
[13] Guimera R, Nunes Amaral L A. Functional cartography of complex metabolic networks [J]. Nature, 2005, 433(7028): 895-900.
[14] De Meo P, Ferrara E, Fiumara G, et al. Mixing local and global information for community detection in large networks[J]. Journal of Computer and System Sciences, 2014, 80(1): 72-87.
[15] Fortunato S, Barthelemy M. Resolution limit in community detection [J]. Proceedings of the national academy of sciences, 2007, 104(1): 36-41.
[16] Mairisha M, Saptawati G A P. Improved modularity for community detection analysis in weighted graph[C]//2016 4th International Conference on Information and Communication Technology (ICoICT). IEEE, 2016: 1-6.
[17] Chen S, Wang Z Z, Bao M H, et al. Adaptive multi-resolution modularity for detecting communities in networks [J]. Physica A: Statistical Mechanics and its Applications, 2018, 491: 591-603.

ISSN 2616-5775 Vol. 5, Issue 12: 82-88, DOI: 10.25236/AJCIS.2022.051212

[18] Qiao Shaojie, Han Nan, Zhang Kaifeng, et al. Overlapping community detection algorithms in complex network big data[J]. Journal of Software, 2017, 28(3):17.
[19] Qiao Shaojie, Guo Jun, Han Nan, et al. Parallel discovery algorithms for large scale complex network communities[J]. Journal of Computer Science, 2017, 40(3): 687-700.
[20] Zhu X, Ghahramani Z. Learning from labeled and unlabeled data with label propagation[R]. City: Citeseer, 2002.
[21] Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks[J]. Physical review E, 2007, 76(3): 036106.