Community Mining Algorithm Based on Structural Similarity

Yaqiong Liu[1], Lu Wang*, Guoqing Chen[2]

[1]College of Information Science and Engineering, Shandong Agricultural University, Tai'an, China
[2]College of Agricultural, Shandong Agricultural University, Tai'an, China

Abstract. In order to improve the efficiency of community mining algorithm and the accuracy of community classification, a community mining algorithm based on structural similarity is proposed in this paper. The algorithm uses the structural similarity as an edge weight to perform the operation of the loop deletion, and implements community merging for isolated nodes, thus improving the precision of community division. The algorithm is compared with GN and SSNCA algorithm in classic data sets such as Zachary network, football data and dolphin social network. The experimental results show that the algorithm can effectively detect the community structure in complex networks, and the accuracy of classification and operation speed are obviously improved.

1 Introduction

In the real world, many systems can be described as networks, such as social networks, biological networks. These networks have high complexity and are therefore called "complex networks" ([1]). Complex networks have become one of the most important interdisciplinary fields of interdisciplinary research. The nodes in the network and the edges represent some kind of connection. In these networks, there are characteristics of the community structure that "the same group of nodes with close connection and sparse connection between different groups" ([2]). The purpose of complex network community mining is to detect and reveal the community structure inherent in heterogeneous complex networks.

At present, domestic and foreign scholars have proposed a variety of community mining algorithms from different perspectives on the issue of community mining. Girvan and others first proposed the GN algorithm ([3]) for community mining. Liu Dayou used structural similarity to replace the edge number in GN algorithm to realize community mining ([4]).Li Zhaonan proposed a community mining algorithm based on the similarity degree of node distance ([5]).For sparse networks, the precision of community division of the above algorithm is not high and the time complexity of the algorithm is high. Based on the above problems, this paper starts with improving the accuracy and operational efficiency of community division presents a community mining algorithm based on structural similarity.

2 Network representation method and formula construction

2.1 Complex Network Diagram Structure

We use a graph G=(V, E)to abstract the complex network. The node V=$\{V_i\}$set represents the individual in the complex network, and the set of edges $E = \{(i,j)|i,j \in V\}$is the relation between individuals.

2.2 Community structure

In complex networks the nodes of the same group are closely linked, and the node links between different groups are sparse and the community characteristics shown are called community structures. In a complex network G, a community is defined as C_i, and the collection of communities is represented as P={C_1,C_2,C_3…C_k}

2.3 Modularity function Q

In 2004 Newman proposed a quantitative standard for characterizing the strengths and weaknesses of the network community structure, which is called the modular function Q ([6]). The community structure P of high module value indicates that the nodes in the community are closely connected and the nodes in the community are sparse. The formal definition of module degree is defined as:

$$Q = \frac{1}{2m}\sum_{ij}((A_{ij} - \frac{d_i d_j}{2m}) \times \partial(v(i), v(j))) \quad (1)$$

$A = (A_{ij})_{n\times n}$ represents the adjacency matrix of the network, If there is an edge connection between node iand node j, then $A_{ij} = 1$,Unbounded connected with

Corresponding author: liuyaq@sdau.edu.cn wangl@sdau.edu.cn

$A_{ij} = 0$. For the function $\partial(u,r)$, if u = r is 1, otherwise, it is 0. The symbol of d_i represents the degree of node i, defined as $d_i = \sum_j A_{ij}$.The symbol of $m = \frac{1}{2}\sum_{ij} A_{ij}$ represents the total number of edges in network G.

3 Data clustering algorithm SSNCA based on K neighbor network

3.1 Introduction of SSNCA algorithm

SSNCA algorithm is a network clustering algorithm based on structured similarity ([7]).The specific steps are as follows:

Step1：Calculate the structural similarity values of all edges in the network.

Step2: Remove the edge with the lowest structural similarity value from the network.

Step3: Recalculate the structured similarity values of all edges affected by step 2.

Step4: Repeat the above steps until there is no edge deletion.

Step5: Select the clustering result that maximizes the value of the Q function as the best community for the network.

3.2 Structure similarity formula construction

Definition 1: The node neighborhood set $\Gamma(i)$ is expressed as the neighbor node of the node i plus its own set.

$$\Gamma(i) = \{j \in V | (i,j) \in E\} \cup \{i\} \quad (2)$$

Definition 2: Let i, j ∈ V, Structural similarity $\sigma(i,j)$ is represented as

$$\sigma(i,j) = \frac{|\Gamma(i) \cap |\Gamma(j)|}{\sqrt{|\Gamma(i)|\Gamma(j)|}} \quad (3)$$

The structural similarity is used to judge the similarity between two points. The larger the value is, the closer the two nodes are, the more likely it is to belong to a community. The SSNCA algorithm uses structural similarity to replace the number of edge in the GN algorithm and improves the calculation speed. However, because only one edge with the smallest structural similarity is deleted at the same time, the computation efficiency is slow, and the problem of isolated points after community division is not considered. Therefore, the effect of community division is not accurate enough, and the calculation efficiency needs to be improved.

4 Community Mining Algorithm SSMCA Based on Structural Similarity

4.1 Introduction of SSMCA algorithm

In order to solve the problem of isolated points in community division, a community mining algorithm SSMCA (the SSNCA based on Merge, SSMCA) based on structured similarity is proposed to improve community accuracy.

Based on the SSNCA algorithm, the algorithm starts from how to accurately divide communities, and adopts circular edge deletion operation to reduce the number of iterations, and then merges the isolated points after dividing the community, so as to achieve the purpose of better division of community structure.

Therefore, this paper introduces the critical value parameter a, used to delete the similarity degree is less than a, in order to avoid blindly selecting the a value and reduce the community classification accuracy, this paper gives the optimal selection strategy: community setting step length in Δa, calculating the module function Q corresponding to the discrete critical value within the critical value interval, find out the critical value corresponding to the maximum Q value and divide the community as the best community division.

4.2 Community similarity formula construction

Definition 3(Community structure similarity) Let i, j ∈ V, community structure similarity C_{ij} expressed as:

$$C_{ij} = \frac{The\ number\ of\ links\ between\ C_i\ and\ C_j}{\sqrt{d_{ci}d_{cj}}} \quad (4)$$

$$d_{ci} = \sum_{j=1}^{|C_i|} degree(V_j) \quad (5)$$

The degree (V_i) indicates the degree of nodes in the graph G, that is, the number of adjacent nodes of V_i.

4.3 SSMCA algorithm steps

Step1：Given the critical interval [x, y], (0<x<y<1), and the step length Δa, the number of cycles in the critical value interval is i, $i = 1$, and $a1 = x + \Delta a$

Step2：The edge is used to calculate the structural similarity between the two nodes connected by the edge, and then the structure similarity is given as the weight to the edge.

Step3：According to the set critical value value a, perform the loop delete edge operation; repeat the steps 2 until no edge is deleted, take the breadth first traversal to obtain the community division under the critical value, then use formula 1 to calculate modularity function Q.

Step4：If $i + 1 < \frac{y-x}{\Delta a}$, then let $a_{i+1} = a_i + \Delta a$, continue to step 3, otherwise stop the loop.Find the community partition of the above maximum Q value as the result of optimal community segmentation.

Step5：The isolated point is considered as an independent community, calculating the similarity of the community structures between them (formula 4, 5) and sorting and merging.

Figure 1 shows a variation curve of the modularity function Q with the critical value a ($\Delta a = 0.02$) obtained by the SSMCA algorithm in the Zachary network. Where the X-axis represents the different

critical value a, and the Y-axis represents the Q value obtained at the selected critical value a. From the figure, we can see that when $a \in [0.3,0.34]$, the maximum Q value is 0.235, which corresponds to the best community division. In the critical value interval [0.3,0.34] the Q value remains unchanged, indicating that deleting a certain part of the network does not affect the overall community structure.

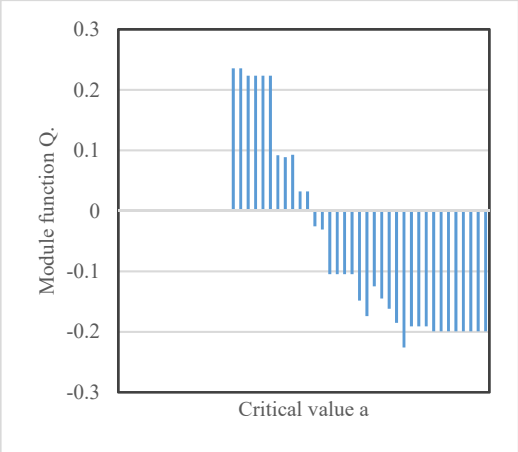

Fig. 1. Zachary network modularity function Q change diagram with critical value a

5 Experimental results and analysis of SSMCA algorithm

This paper uses SSMCA algorithm to perform simulation experiments on classical data sets. The results are compared with GN algorithm and SSNCA algorithm. Experiments show that the SSMCA algorithm can effectively detect the community structure effectively, and the calculation speed and community division accuracy are significantly improved.

The experimental environment of this algorithm is: Intel(R) Core(TM) i5-6600 CPU@3.30GHZ, memory 4GB, operating system Windows 7 Ultimate Edition, and programming software MATLAB 2014b.

5.1 Comparison of SSMCA algorithm and classical algorithm results

5.1.1 Zachary network

Figure 2 shows the Zachary network ([8]) nodes 1 and 33 in the figure represent the club managers and the coaches. The nodes of different shapes and colors represent members of each group after splitting.

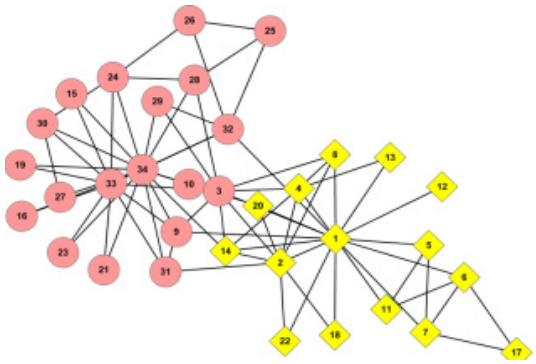

Fig. 2. Original network

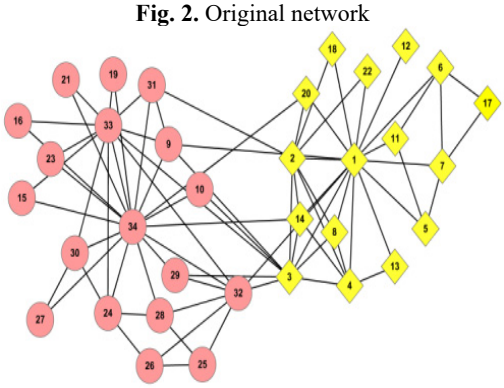

Fig. 3. GN algorithm detection network

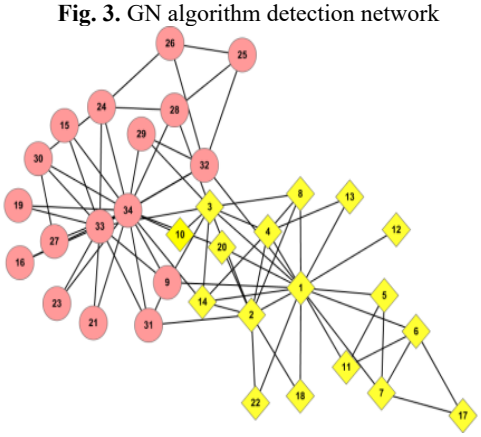

Fig. 4. SSNCA algorithm detection network

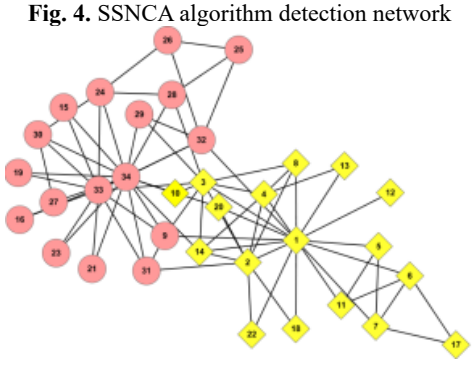

Fig. 5. SSMCA algorithm detection network

Using the GN algorithm, the SSNCA algorithm obtains Figure 3 and Figure 4. The result of the community division using the SSMCA algorithm is shown in Figure 5. In the figure, the yellow diamond node set and the red circular node set represent two

divided communities respectively. SSMCA algorithm and SSNCA algorithm are almost identical (see Figure 4 for details). After investigation and comparison, the new algorithm is consistent with the real situation.

5.1.2 Dolphin social network data set

The dolphin data set was obtained by D. Lussean includes 62 nodes and 159 edges ([9]). As shown in Figure 6, the circle in the figure represents the node set that successfully belongs to the community, and the different colors represent different communities. Unfortunately, there are isolated points 37, 40, and 56 (shown as triangles) after the community was divided into communities.

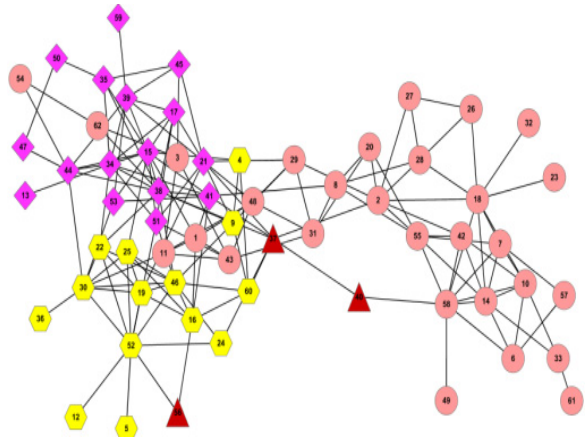

Fig. 6. SSNCA algorithm detection network

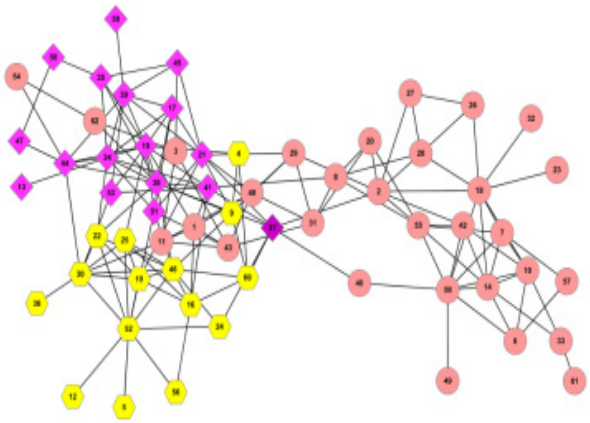

Fig. 7. SSMCA algorithm detection network

Using our new algorithm, SSMCA, the community division was mapped to figure 7, and the new algorithm also divided the network into three communities, with three different shapes of points representing different three communities. Unlike SSNCA algorithm in this paper, the algorithm of the isolated points in 37, 40, 56 in the community, to achieve the purpose of the new algorithm is better than SSNCA algorithm, and the experimental results are basically consistent with the actual network community.

5.1.3 American university football data set

The U.S. college football data set contains a total of 115 nodes and there are 613 connected edges. ([10]).The SSMCA algorithm was used to divide the community. The results are shown in Figure 8. The algorithm divides the network into 12 communities, and Figure 9 gives a statistical graph of the number of teams in each community, which is basically between 7 and 11 branches, which is consistent with the actual situation Therefore, the algorithm can find the community structure effectively.

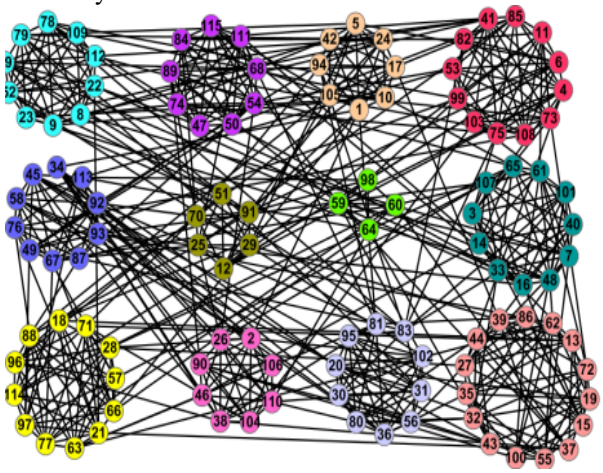

Fig. 8. Football Community Divisio

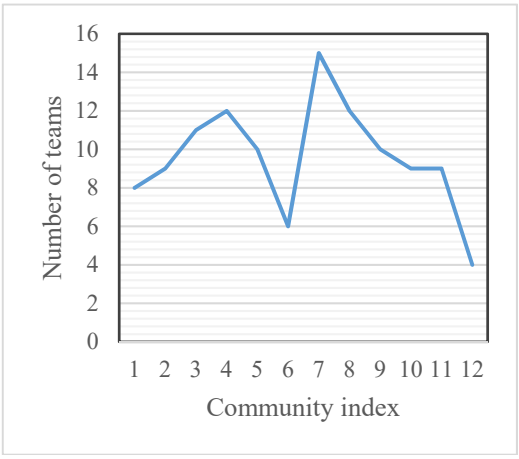

Fig. 9. Community Team Statistics

5.2 Comprehensive comparison of data sets

Experiments show that the SSMCA algorithm has significantly improved the accuracy of community division, and effective community consolidation has been implemented for the isolated points that existed after the community was divided, and the operation efficiency of the algorithm has been greatly improved.

Table 1 shows the statistical data of the experimental network. Table 2 shows the time consuming and accuracy standard comparison of GN algorithm, SSNCA algorithm and SSMCA algorithm for community partition of Zachary data set, the dolphin social network, and the

football data set. We use the metric system based on Information theory NMI as the accuracy measure standard of the algorithm ([11]) The values in the table are the average values obtained 100 times.

For different data sets of the same algorithm, the larger the network size, the longer the operation time. For the same data sets of different algorithms, our algorithm can reasonably divide the network in terms of accuracy standards, and the running time is the fastest when the length is set to 0.02, followed by SSNCA, and the GN algorithm is the slowest. In terms of accuracy, the SSNCA algorithm in the dolphin social network has three isolated points. The SSMCA algorithm successfully classifies outliers into the community, which solves the problem well. In the football data set, the GN algorithm only divides 11 communities. SSMCA algorithm is well divided into 12 communities that match the actual results. In conclusion, the experimental results show that the proposed algorithm can effectively find the community structure effectively and significantly improve the computational efficiency and community classification accuracy of the SSNCA algorithm.

Table 1. Data Set Statistics

Data set	Node	edge	The network diameter
Zachary Network	34	78	5
Dolphin social network	62	159	8
Football data set	115	613	4

Table 2. Comparison of SSMCA and other algorithms on real networks known for three cluster structures

	Zachary Network		Dolphin social network		Football data set	
algorithm	time	NMI	time	NMI	time	NMI
GN	0.28	57.98 %	1.20	44.17 %	14.84	87.89 %
SSNCA	0.21	68.37 %	1.06	52.95 %	11.04	88.49 %
SSMCA	0.19	89.54 %	0.57	87.66 %	1.796	92.68 %

6 Conclusion

This paper aims at improving the SSNCA algorithm, using the community similarity formula to merge the isolated points that exist after the community division, and on this basis, a community similarity mining algorithm SSMCA based on structural similarity is proposed. The algorithm can solve the problem of isolated points left after community division, and it can improve the accuracy of community division, and greatly shorten the running time of algorithm greatly. The algorithm is applied to real data set Zachary data set, dolphin social network, football data set, and compared with GN and SSNCA algorithms. Experiments show that the algorithm has obvious advantages in community division accuracy and running time. The next step will be to improve the accuracy of the algorithm for community segmentation of large networks.

Acknowledgment

National Natural Science Foundation of China (91746104) The Optimum Allocation of Annual Light and Temperature Resources and Group Regulation Techniques for Wheat and Maize (005-35432)

References

1. Newman M E J, Detecting community structure in networks[J].European Physical Journal B,**38**（2）：321-330(2004).
2. Berger-Wolf T Y, Saia J. A framework for analysis of dynamic social networks[C]// Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, Pa, Usa, August. DBLP:523-528,(2006).
3. Girvan M, Newman M E J, Community structure in social and biological networks [J]. Proceedings of the National Academy of Sciences of the United States of America, **99**(12):7821-6(2002).
4. Bo Y, Liu D Y, Liu J, et al. Complex Network Clustering Algorithms[J]. Journal of Software, **20**(1) (2009).
5. Zhaonan L I, Yang B, Liu D. Distance Similarity Algorithm for Mining Communities from Complex Networks[J]. Journal of Frontiers of Computer Science & Technology, **5**(4):336-346(2011).
6. Newman M E J, Girvan M. Finding and evaluating community structure in networks [J]. Phys. Rev. E, **69**(2):026113(2004).
7. Kim brother, Liu Jie, Jia Zhengxue, Liu Dayou, Data clustering algorithm of k nearest neighbor network based on [J].Pattern recognition and artificial intelligence, **23** (4): 546-551(2010).
8. Zachary W W. An Information Flow Model for Conflict and Fission in Small Groups[J]. Journal of Anthropological Research, **33**(4):452-473(1977).
9. Lusseau D. The emergent properties of a dolphin social network.[J]. Proceedings Biological Sciences, **270** Suppl 2(supplement 2):S186(2003).
10. Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proceedings of National Academy of Science, **9**(12): 7821-7826(2002).

11. Danon L, Duch J, Diaz-Guilera A, et al. Comparing community structure identification. J Stat Mech: P09008[J]. Journal of Statistical Mechanics Theory & Experiment, 2005(9) (2005).