Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 3, No. 3, 2023

Multi-objective Optimization Overlapping Community Detection Algorithm based on Subgraph Structure

Changhong Li *

School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan Hubei, 430074, China

Abstract: Community detection in complex networks has increasingly become an important topic in the network, but in most community detection methods, a single node only belongs to one community. In fact, there is often overlap among real-world online communities. In this paper, an overlapping community detection algorithm based on subgraph structure and multi-optimization method is designed. In this algorithm, the maximum clique mined by k-core decomposition is used as the clique node, thus the overlap characteristic is transformed into the inherent characteristic of the new graph. After that, a population initialization method based on k-core decomposition is designed, and the discrete framework of multi-objective particle swarm optimization algorithm is used to optimize the two objectives on the basis of maximal clique graph to solve the problem of overlapping community detection. In the real-world network, this algorithm is compared with similar community detection algorithms. The comparison of the evaluation indexe shows that the community detection effect of this algorithm is similar to that of similar algorithms, and has a good application prospect.

Keywords: Overlapping Community Detection; Evolutionary Algorithm; Multi-objective Optimization.

1. Introduction

In recent years, more and more network technology researchers have modeled the real world as a complex network. Typical real-world complex networks include the World wide Web, neural networks[1], biological networks[2] social networks[3] and so on. These complex networks have profound research value, and people hope to provide solutions to real-life problems through the study of complex networks, for example, mining the key gene modules of cancer, detecting potential terrorists through interpersonal networks, simulating the complex functions of biological networks to solve problems, and so on. Researchers usually represent the individuals in the network as nodes, and the relationships between individuals as the edges of connected node[4].

In a complex network, individuals gather together to form a community. Network community structure detection, also known as network clustering, is a technology to reveal the community structure formed by network aggregation. Community detection divides the network into several societies, which is based on the fact that the edge connections between nodes in the same community should be tight, while the edge connections between inter-community nodes should be sparse[5]. Generally speaking, the community structure can be regarded as a collection of interconnected information carriers with the same or similar characteristics of some owners. Community detection is helpful for us to study groups with more similarities, infer the future trend of research objects, and predict the nature or relationship between groups or individuals.

Community detection is gradually developed from the early graph division and hierarchical clustering. The analysis and research of community structure has a long history, and has evolved a variety of different algorithms. Newman et al proposed the concept of community for the first time and gave a top-down hierarchical clustering method GN algorithm. Subsequently, experts in different fields have proposed a lot of community detection algorithms, these algorithms can be divided into the following categories: methods based on hierarchical structure classification, methods based on modularity optimization, spectral clustering methods, random walk and evolutionary computing methods.

In the traditional network clustering research, the above community detection methods are mostly separated community detection without additional representation, that is, a node in the graph can only belong to one community at the same time[6]. In the real world, however, clusters often overlap[4] [7]. Therefore, the methods of overlapping community detection continue to appear in recent years.

Because of the particularity of the structure, the algorithm derived from overlapping community detection is quite different from that of separated community detection. The main methods of overlapping community detection include algorithm based on factional filtering, algorithm based on local extension, algorithm based on edge partition, algorithm based on fuzzy detection, algorithm based on k-clique seepage, overlapping community detection algorithm based on local expansion and optimization, and overlapping community detection algorithm based on multi-objective optimization.

Among a variety of algorithms, the multi-objective optimization algorithm has a strong advantage in solving the problem of overlapping community detection. Because community detection can be regarded as the objective optimization problem of the density of inter-community links in the community kernel, and this kind of problem is often a NP difficult problem, which is difficult to be solved by ordinary algorithms. Therefore, many scholars introduce the multi-objective optimization algorithm suitable for solving this kind of problem into the overlapping community detection problem, such as MCMOEA[4], MR-MOEA[7] and so on, and achieved good results. Scholars generally believe that societies should have dense internal connections and sparse and inter-group connections, which means that community detection needs to optimize two contradictory goals, namely, maximize internal links and minimize external links[4]. Therefore, the community detection problem is transformed into a multi-objective optimization problem,

with these two objectives as the core, there are specific objectives of different representation methods, among which kernel k-means (KKM)[8] and ratio cut (RC)[9] are derived from machine learning algorithms.

2. Methodology

2.1. Construction of Maximal Clique Graph based on K-core Decomposition

K-core is a subgraph of a graph G, where the degree of each vertex in the subgraph is at least k. While k-clique is that k nodes are connected to each other, then each node must have a degree of at least k-1, so kclique must be a subgraph of (k-1)-core, so we can search for the maximum k-clique of each node in the subgraph of k-core decomposition. As shown in Figure 1, the degree of the 11 nodes that make up the network is greater than or equal to 2, so the whole network is 2-core. Among them, the nodes {1, 2, 3, 4, 5, 6, 7, 8} constitute a 3-core, while the node set {1, 2, 3, 4} and {5, 6, 7, 8} are two 4-clique.

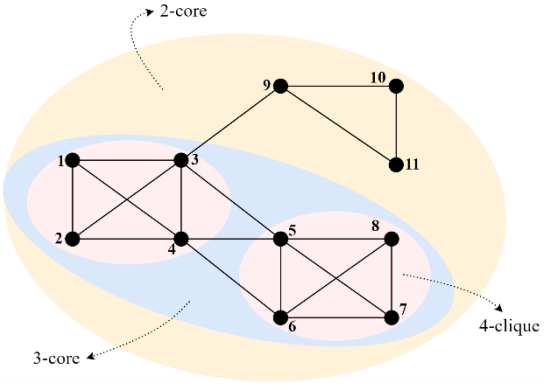

Figure 1. Schematic diagram of the relationship between K-core and K-clique

Therefore, the k-core decomposition graph is adopted, and the k values of all k-core in the graph are arranged from large to small. For each k-core subgraph, the maximal clique graph search strategy is used to find the node subgraph[4]. The maximum clique problem of each node is a complete NP problem, so the subgraph decomposed by k-core can reduce the node dimension and speed up the search speed. Moreover, when the k is large, although the amount of calculation of the connected edge is large, the number of nodes is small. In the case of k=1, although the subgraph is the original graph and the number of nodes is not reduced, it is only necessary to calculate the 2-clique of previously unassigned nodes, which is less difficult to calculate. Generally speaking, the maximal clique graph mining method based on k-core decomposition has relatively stable performance.

After obtaining the maximum clique to which each node belongs, considering the overlapping nodes, overlapping edges and connected edges in the clique nodes, the edges of the maximal clique graph are constructed, and after filtering out the noise, the maximal clique graph with maximal clique nodes is obtained[4]. In the edge noise filtering method, the threshold which can get the maximum modularity is selected. The node of a maximal clique graph itself contains multiple nodes, and a single original node can be subordinate to multiple maximal cliques, which completes the overlapping representation of nodes.

2.2. Multi-objective Optimization Framework

In traditional algorithms, many initializations are random initialization, that is, random numbers are generated, which leads to the confusion of the whole network division at the beginning stage, the lack of diversity of the population, and the low quality of each solution, which slows down the speed of calculation. The closer it is to the initialization of the real network partition, the better the performance of the algorithm. A commonly used population initialization method is label propagation, which assigns the largest number of tags in neighbors to each node. However, the existing label propagation algorithms ignore the attributes of the network itself.

K-core decomposition can roughly divide the community and can be used as the initialization algorithm of the population. In the label propagation algorithm, the update order of nodes significantly affects the formation of the community, and strengthening the propagation priority of the nodes with lower node degree can improve the stability and balance of the community structure. Based on the label propagation method, this paper sets the label of the cluster node belonging to the largest k-core to the same tag, and adds the balance strategy to realize the population initialization[10].

This paper adopts the multi-objective optimization framework and runs the MODPSO algorithm according to the partition results of the maximal clique graph obtained by the initialization algorithm. MODPSO is a particle swarm optimization algorithm based on decomposition. This method has fast convergence speed, high accuracy of community division results, and is suitable for weight networks.

3. Results and Discussion

3.1. Evaluation Index

Each round of operation of the multi-objective algorithm can produce a set of non-dominant solutions. The two goals set by the algorithm do not have the problem of which is superior or inferior, and the result is a compromise solution set. The algorithm needs to introduce evaluation index to judge the partition. In this paper, generalized normalized mutual information (gNMI)[11] is used as the evaluation standard.

The result of the calculated partition and the gNMI value of the real partition can evaluate the proximity between the result of the community detection algorithm and the real partition. The larger the value of gNMI is, the closer the community structure is to the real partition.

3.2. Real Network Experiments

In all experiments, the population size and maximum algebra were set to 100 recommended in MODPSO. All the experiments were carried out on Intel Core i5-8265U (1.60 GHz, 1.80 GHz) CPU, 8 GB RAM and Windows 10 computers based on x64 processors.

This article refers to five real online data sets, which are Zachary's karate network[12], dolphin social network[13], American college football network and American political books network [14]. All four real networks have real community divisions, so gNMI can be used to measure the effectiveness of the four network divisions. Table 1 shows the basic information of these four networks.

Table 1. Structural characteristics of four Real Networks

Network	Number of nodes	Number of edges	Average degree	Number of real communities
football	115	613	10.66	12
karate	34	78	4.59	2
dolphin	62	159	5.13	2
polbook	105	441	8.4	3

Table 2. Comparison of algorithms on Real Network Maximum gNMI Indexes

Nework	This article	MCMOEA	MODPSO	MR-MOEA	NMF	IMOQPSO	+/-/≈
football	0.803	0.712+	0.9289-	0.803	0.793+	0.809	2/1/2
karate	0.918	0.918	1-	1-	0.837+	0.708+	2/2/1
dolphin	0.889	0.473+	1-	1-	0.907-	1-	1/4/0
polbook	0.447	0.104+	——	0.149+	0.388+	0.432+	4/0/0

+ indicates that the performance of this algorithm in obtaining the maximum gNMI partition is better than that of other methods; -indicates that the performance of this algorithm in obtaining the maximum gNMI partition is weaker than that of other methods; ≈ indicates that the performance of this algorithm is similar to that of other algorithms in obtaining the maximum gNMI partition.

Table 2 shows the maximum gNMI comparison results of the proposed algorithm and the other five algorithms applied to the above four real networks. Generally speaking, the experimental results of this algorithm on these four real networks are competitive.

4. Conclusion

In this paper, we propose a multi-objective optimal overlapping community detection algorithm based on subgraph structure, which detects the maximum clique in the network based on k-core decomposition, and transforms the maximal clique into clique nodes to construct a weighted network. Then the initialization method of tag propagation population based on balance and k-core decomposition is introduced, and then the multi-objective particle swarm optimization algorithm is used to complete the community detection of the network. The experimental results show that the algorithm proposed in this paper achieves good results in both synthetic network and real network.

References

[1] Watts, D., Strogatz, S. Collective dynamics of ‘small-world’ networks. Nature 393, 440–442 (1998).

[2] Pizzuti C, Rombo S E. Algorithms and tools for protein-protein interaction networks clustering, with a special focus on population-based stochastic methods. 2014.

[3] Wasserman, Stanley, Faust, et al. Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)[M]. Cambridge University Press, 1994.

[4] Wen X, Chen W N , Ying L , et al. A Maximal Clique Based Multiobjective Evolutionary Algorithm for Overlapping Community Detection[J]. IEEE Transactions on Evolutionary Computation, 2017, PP (99):1-1.

[5] Girvan M, Newman M E . Community structure in social and biological networks[J]. Proc Natl Acad, U S A, 2002, 99 (12): 7821-7826.

[6] Gong M, Cai Q , Chen X , et al. Complex Network Clustering by Multiobjective Discrete Particle Swarm Optimization Based on Decomposition [J]. IEEE Transactions on Evolutionary Computation, 2014, 18(1):82-97.

[7] Zhang L, Pan H , Su Y , et al. A Mixed Representation-Based Multiobjective Evolutionary Algorithm for Overlapping Community Detection[J]. IEEE Trans Cybern, 2017:1-14.

[8] Pizzuti C. GA-Net: A Genetic Algorithm for Community Detection in Social Networks[J]. Springer Berlin Heidelberg, 2008.

[9] Angelini L Boccaletti S , Marinazzo D , et al. Identification of network modules by optimization of ratio association[J]. Chaos: An Interdisciplinary Journal of Nonlinear Science, 2007, 17(2):023114-.

[10] H. Roghani and A. Bouyer, "A Fast Local Balanced Label Diffusion Algorithm for Community Detection in Social Networks," in IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 6, pp. 5472-5484, 1 June 2023, doi: 10.1109/TKDE.2022.3162161.

[11] Huawei Shen, Xueqi Cheng, Kai Cai, et al. Detect overlapping and hierarchical community structure in networks[J]. Physica A, 2009, 388(8):1706-1712.

[12] Wayne, W, Zachary. An Information Flow Model for Conflict and Fission in Small Groups[J]. Journal of Anthropological Research, 1977, 33(4):452-473.

[13] Lusseau D. The emergent properties of a dolphin social network[J]. Proceedings of the Royal Society B: Biological Sciences, 2003.

[14] M. E. J. Newman. Modularity and community structure in networks[J]. Proc. Nat. Acad. Sci. USA, vol. 103, no. 23, pp. 8577–8582, 2006.