Locally Expanded Genetic Optimization Method for Overlapping Community Detection (Post-print)

Authors: Chu Yangjie, Yang Zhongbao, Hong Ye

Date: 2018-05-18T00:00:00+00:00

Abstract

Overlapping community structure constitutes an important characteristic of complex networks. This paper proposes a Locally Expanding Genetic Optimization for Overlapping Community Detection (LEGAOCD). Drawing upon the concept of locally expanding overlapping community detection methods, motifs are constructed from a small number of core nodes. Simultaneously, triangular motifs are utilized to address the community stability measurement problem, thereby quantifying the stability of community structure. Subsequently, an improved genetic optimization algorithm strategy is employed to assign them to their appropriate communities. Finally, high-quality overlapping community structures are obtained through two evaluation objective functions. The algorithm is compared with the classic CPM algorithm and COPRA algorithm on datasets, and experimental results demonstrate that the LEGAOCD algorithm exhibits superior performance in detecting overlapping community structures and overlapping nodes.

Full Text

A Local Extension Approach through Genetic Algorithm for Overlapping Community Detection

Chu Yangjie, Yang Zhongbao, Hong Ye
(School of Science, Wuhan University of Technology, Wuhan 430070, China)

Abstract: Overlapping community structure represents a crucial characteristic of complex networks. This study proposes a Local Extension Genetic Algorithm Optimization for Overlapping Community Detection (LEGAOCD). Drawing inspiration from local extension methods for overlapping community detection, the algorithm constructs motifs from a small number of core nodes; simultaneously, it utilizes triangular motifs to assess community stability measures,

thereby quantifying community structure stability. Through an improved genetic optimization algorithm strategy, it then allocates appropriate communities for these nodes. Finally, high-quality overlapping community structures are obtained via two evaluation objective functions. Compared with classical CPM and COPRA algorithms on benchmark datasets, experimental results demonstrate that LEGAOCD achieves superior performance in detecting overlapping community structures and identifying overlapping nodes.

Keywords: local extension; genetic algorithm; overlapping community detection; core node; multi-objective optimization

0 Introduction

In static network environments, overlapping community detection algorithms can be broadly categorized into five classes: clique percolation methods, local extension methods, line graph partitioning methods, fuzzy clustering methods, and agent-based methods [1]. Among these, classic algorithms include the Clique Percolation Method (CPM) [3], the COPRA algorithm based on agent methods [4], and the Speaker-Listener Label Propagation Algorithm (SLPA) [5]. The classical clique percolation algorithm belongs to the clique filtering approach, which can discover larger communities but shows low probability in mining small communities, with excessive time and space overhead that prevents its application to large-scale networks. The COPRA algorithm, based on agent methods, randomly selects labels, leading to poor convergence performance and unstable overlapping community division results. SLPA, also an agent-based method, primarily sets different probability criteria and adjusts parameters for different labels, making parameter tuning difficult for different networks in practical applications and hindering generalization.

This paper proposes a Local Extension Genetic Algorithm Optimization for Overlapping Community Detection (LEGAOCD). Inspired by local extension methods for overlapping community detection, the algorithm constructs motifs from a small number of core nodes. Simultaneously, it employs triangular motifs to evaluate community stability measures, thereby quantifying community structure stability. Through an improved genetic optimization algorithm strategy, it allocates appropriate communities for these nodes. Finally, high-quality overlapping community structures are obtained via two evaluation objective functions. Experimental comparisons with classical CPM and COPRA algorithms on benchmark datasets demonstrate that LEGAOCD exhibits superior performance in detecting overlapping community structures and identifying overlapping nodes.

1 Basic Concepts

Definition 1 (Overlapping Network Community Structure). Network community structure is a partition scheme $\mathcal{P} = \{c_1, c_2, \cdots, c_m\}$ of the network node set, where each community c_i must satisfy $c_i \subseteq V$, $c_i \neq \phi$ for $i = 1, 2, \cdots, m$, and $\bigcup_{i=1}^{m} c_i = V$. For any two communities c_i and c_j, if $c_i \cap c_j \neq \phi$ and $i \neq j$, then $\mathcal{P}$ is called an overlapping community structure.

Definition 2 (Maximum Core Node Degree). Let the label set of core nodes be $\Omega = \{u_1, u_2, \cdots, u_k\}$. The maximum core node degree is defined as $N(u_t) = \text{argmax}_{u \in \Omega}(p(u) \cdot \text{ion}(u))$, where $\deg(u_t) = k$ represents the degree of node u_t, and $\text{ts}(u_t) = m$ represents the strength of node u_t.

Definition 3 (Common Neighbors of Core Nodes). The common neighbor between core node u_i and node u_j is defined as $C_{ij} = \text{neighbor}(u_i) \cap \text{neighbor}(u_j)$, where $\text{neighbor}(u_i)$ denotes the neighbor set of core node u_i.

Definition 4 (Motif Structure). A motif exists between individuals and communities, composed of connections among a few nodes. It represents the basic connection pattern among members within a community, revealing the evolutionary 规律 of networks.

Definition 5 (Weighted Community Clustering). Weighted Community Clustering (WCC) [8] is an objective function for measuring network topological properties. It depends on triangular motifs (three-node motifs) within communities to determine community stability, thereby quantifying community structure quality. The weighted community clustering is defined as:

$$WCC(u, NP) = \begin{cases} 0 & \text{if } NP = \phi \\ \frac{(\sum_{v \in NP} t_{uv}) + t_{uv}}{NP(NP-1)} & \text{if } \exists v \in NP, t_{uv} > 0 \\ 0 & \text{otherwise} \end{cases}$$

where t_{uv} represents the relationship coefficient between nodes u and v; $t_{uv} > 0$ indicates a positive correlation, while $t_{uv} < 0$ indicates a negative correlation. In an unweighted network $G = (V, E)$, if nodes i and j are connected by an edge, then $e_{ij} = 1$; otherwise, $e_{ij} = 0$. Therefore, the degree of node i is $k_i = \sum_{j \in V} e_{ij}$, representing the number of edges connected to node i. The node weight s_i is defined as the sum of weights of edges associated with it, also called node strength.

The relationship between community node strength and WCC is:

$$F(u, c) = \frac{s_{c_u}^{in+} + s_{c_u}^{in-}}{s_{c_u}^{in+} + s_{c_u}^{in-} + s_{c_u}^{out+} + s_{c_u}^{out-}}$$

where $s_{c_u}^{in+}$ denotes the positive internal node strength within community c; $s_{c_u}^{in-}$ denotes the negative internal node strength within community c; $s_{c_u}^{out+}$ denotes

the positive external node strength within community c; and $s_{c_u}^{out-}$ denotes the negative external node strength within community c.

Definition 6 (Node-Community Closeness). Based on literature [9], the closeness between a node and a community is defined as:

$$OD(u,c) = \frac{a_{c_u}^{in+} + a_{c_u}^{in-}}{a_{c_u}^{in+} + a_{c_u}^{in-} + a_{c_u}^{out+} + a_{c_u}^{out-}}$$

where $a_{c_u}^{in+}$ represents the number of nodes in community c that form triangular motifs with node u and node set c; $a_{c_u}^{in-}$ represents the number of nodes in node set c that form at least one triangular motif with node u; and $|c|$ represents the number of nodes in set c excluding node u.

The closeness between overlapping nodes and communities is defined as:

$$F(u,c) = \frac{1}{|c|}\sum_{i=1}^{|c|} OD(u,c_i) \quad \text{for } u \in c$$

Definition 7 (Modularity). Modularity quantitatively evaluates the overall quality of overlapping community partitions. Based on positive and negative correlation edge weights, an appropriate modularity measure is adopted. According to literature [9], the improved definition is:

$$Q = \sum_{c\in\mathcal{P}} \left[\frac{s_c^+}{s_c^+ + s_c^-} - \left(\frac{2s_c^+ + s_c^{out+}}{2s_c^+ + 2s_c^- + s_c^{out+} + s_c^{out-}} \right)^2 \right]$$

where s_i^+ and s_i^- represent the sum of all positive and negative correlation weights for node i, respectively; specifically, w_{ij} represents the adjacency matrix of network correlation weights. If node i belongs to community c, then $\delta_{ic} = 1$; otherwise, $\delta_{ic} = 0$.

---

2 Local Extension Genetic Algorithm Optimization for Overlapping Community Detection

This paper addresses the overlapping community detection problem in networks using a genetic optimization algorithm framework, aiming to discover high-quality overlapping community structures. The proposed algorithm flow is as follows:

Algorithm 1: LEGAOCD Algorithm

Input: Network topology structure $G = (V, E, W)$
Output: Overlapping community partition set $\mathcal{P} = \{c_1, c_2, \cdots, c_m\}$

1. Encode core nodes of the network;
2. Construct adjacency matrix A and node strength correlation matrix W for core nodes;
3. Initialize population;
4. **while** termination condition is not met **do**
5. **for each** core node $u \in V$ **do**
6. `Initialize $NP_0=\{u\}$;`
7. `**while** $WCC(u,NP_t)\neq 0$ and $n_k\leq |V|$ **do**`
8. `  $t\leftarrow t+1$;`
9. `  **if** $WCC(u,NP_t\cup\{v\})\geq WCC(u,NP_t)$ **then**`
10. `   $NP_{t+1}\leftarrow NP_t\cup\{v\}$;`
11. `**end if**`
12. **end while**
13. Calculate fitness function $F(u, c)$; Selection operator;
14. Evolutionary operations, including uniform crossover and mutation;
15. Offspring population, merge populations, elite selection, return to step 3;
16. **end for**
17. **end while**
18. **foreach** overlapping node **do**
19. **if** $WCC(u, NP_t \cup \{v\}) \geq WCC(u, NP_t)$ **then**
20. $NP_{t+1} \leftarrow NP_t \cup \{v\}$;
21. **else**
22. Merge community structures, similar to algorithm step 9;
23. **end if**
24. **end foreach**
25. Obtain high-quality overlapping community structure;
26. **end**

2.1 Individual Encoding and Decoding

During initialization, all network nodes, communities, and the mapping between nodes and populations, individuals, and genes are established. Drawing inspiration from label propagation algorithms, which can generate individuals with

certain community structures, nodes with greater node strength exert greater stability and influence on their affiliated communities. Individuals are generated randomly, with multiple individuals forming a population. This paper adopts a string encoding scheme, where non-overlapping core nodes belong to only one community identifier, while overlapping core nodes possess multiple identifiers. [Figure 1: see original paper] illustrates a network partition and its corresponding encoding. As shown in the figure, the network may contain two individuals $(1, 1, 12, 2, 2, 2)$ and $(2, 2, 3, 3, 3, 3)$. The first individual contains two overlapping communities $\{1, 2, 3\}$ and $\{3, 4, 5, 6\}$; the second individual contains two non-overlapping communities $\{1, 2\}$ and $\{3, 4, 5, 6\}$.

If these two individuals undergo crossover operation, the result might be $(1, 1, 12, 2, 2, 2)$ and $(1, 1, 12, 2, 2, 2)$, which are then crossed with other individuals to retain community structures with stronger stability. During the decoding process, for any individual, if its initial community label is $(1, 1, 1, 1, 1, 1)$, the entire community structure would be destroyed. The decoding process assigns node label values representing community membership numbers. If at this time $NP(2) = 2$ is the current maximum community label value, the process ends only when $NP(1) = NP(2)$. Throughout the decoding process, individuals $(1, 1, 12, 2, 2, 2)$ and $(2, 2, 3, 3, 3, 3)$ are decoded, and nodes with the same label value are merged into the same community structure.

2.2 Selection

The purpose of selection is to choose excellent individuals from the current population, giving them opportunities to serve as parents for producing offspring in the next generation. Genetic algorithms embody this idea through the selection process, where the principle is that individuals with stronger adaptability have higher probabilities of contributing one or more offspring to the next generation. Selection reflects Darwin' s principle of survival of the fittest. This paper adopts roulette wheel selection [11], using the calculation formula from Definition 6 as the fitness function to obtain high-quality overlapping community structures.

2.3 Genetic Evolution Operation Update Strategy

Genetic evolution operations include crossover and mutation. Crossover produces new offspring by exchanging partial genes between parent chromosomes, with new individuals combining characteristics from their parent individuals. Crossover embodies the idea of information exchange. This paper employs a uniform crossover operator with two-dimensional crossover [12], pairing parent chromosomes pairwise to randomly generate a uniform block crossover operator. The crossover operator exchanges gene portions within rectangular blocks between two parent chromosomes according to a preset crossover probability. Both motifs, individuals, and communities can be represented by rectangular blocks. For example, after crossing two rectangular blocks $(1, 1, 12, 2, 2, 2)$ and $(2, 2, 3, 3, 3, 3)$ and decoding, the result is $(1, 1, 12, 2, 2, 2)$ and $(1, 1, 12, 2, 2, 2)$, thereby changing node label values. The size and position of rectangular blocks

are randomly generated. Using uniform block crossover ensures that each gene in offspring chromosomes corresponds to existing adjacent edges, allowing the algorithm to continue network structure partitioning toward the direction of rectangular block (motif) stability. Mutation randomly selects an individual from the population and, with a certain probability, randomly changes the label value of a node within a community. The mutation operator changes the label value of a node' s community at a preset probability, with mutation occurring at a very low probability.

Algorithm Time Complexity Analysis: Let n be the number of network nodes, k be the number of core nodes in the network, and m be the number of overlapping community structures. The LEGAOCD algorithm' s first two lines take $O(m+k)$ time. Lines 3-16 belong to the genetic process, with algorithm decoding time of $O(k)$, and selection, crossover, and mutation operations taking $O(p)$ time. Lines 17-26 take $O(n)$ time. In the genetic algorithm, population size is p and iteration count is g. Therefore, the time complexity of LEGAOCD is $O(g \cdot p \cdot (m+k+n+p)) = O(g \cdot p \cdot (m+k+n))$.

---

3 Experimental Results and Analysis

LEGAOCD, CPM, and COPRA algorithms were all implemented in Matlab (R2010b). The experimental environment was: Windows 7 operating system, AMD A-8 2.10GHz, 500GB memory. According to genetic algorithm parameter settings, the Normalized Mutual Information (NMI) value does not show significant variation with changes in crossover probability and mutation probability. Generally, high crossover probability and low mutation probability are adopted [13]. Therefore, to facilitate performance testing, experimental parameters were set the same as in a novel genetic algorithm for overlapping community detection [14]: control parameter $a = 0.97$, crossover probability $cp = 0.8$, mutation probability $mp = 0.2$, population size $p = 200$, iteration count $g = 100$. Results for LEGAOCD, CPM, and COPRA algorithms were averaged over 50 runs.

3.1 Evaluation Criteria

F-score is calculated from precision and recall to measure the accuracy of overlapping nodes detected by an algorithm [7]. The formula is derived as follows:

$$P = \frac{|L_{\text{result}} \cap L_{\text{true}}|}{|L_{\text{result}}|}, \quad R = \frac{|L_{\text{result}} \cap L_{\text{true}}|}{|L_{\text{true}}|}, \quad F = \frac{2PR}{P+R}$$

where L_{result} represents communities obtained by the algorithm and L_{true} represents real communities; P is precision, the ratio of correctly detected overlapping nodes to all detected overlapping nodes; R is recall, the ratio of correctly detected overlapping nodes to the actual number of overlapping nodes in the network.

Extended Normalized Mutual Information (NMI) describes the correlation between partition results and true structures [15], with a value range of [0,1]. Values closer to 1 indicate better partition results. The mathematical description is:

$$NMI = \frac{1}{2}\left[\frac{H(Y) - H(Y|X)}{H(Y)} + \frac{H(X) - H(X|Y)}{H(X)}\right]$$

where $H(X|Y)$ represents the normalized conditional entropy of partition X given partition Y.

3.2 Algorithm Comparison and Evaluation

Synthetic datasets primarily consist of LFR benchmark graphs [16]. Different parameters generate different LFR benchmark graphs, creating various types of complex networks. N represents the number of network nodes, $\langle k \rangle$ is the average node degree, $k_{\max}$ is the maximum degree, μ is an adjustable mixing parameter, O_n and O_m represent the number of overlapping nodes and the number of communities each overlapping node belongs to, respectively, τ_1 is the power-law distribution exponent for node degrees, τ_2 is the power-law distribution exponent for community sizes, and $c_{\min}$ and $c_{\max}$ represent the minimum and maximum community sizes in the network. When LFR benchmark network parameters are set as $N = 5000$, $\langle k \rangle = 2$, $k_{\max} = 1$, $c_{\min} = 10$, $c_{\max} = 50$, mixing parameter $\mu = 0.3$ (representing the connection rate between a node and other nodes within the same community), community size range $(20, 100)$, overlapping node count O_n set to 10% of all network nodes, and repeated node community membership count O_m set to $\{2, 3, 4, 5, 6\}$. Larger μ values indicate greater difficulty in partitioning overlapping communities. Basic information for real-world datasets is shown in .

The NMI values of three algorithms on synthetic datasets are affected by mixing parameter μ and overlapping node community membership count O_m. As these increase, partitioning overlapping communities becomes more difficult. As shown in [Figure 2: see original paper], when $\mu = 0.3$ and O_m varies from 2 to 6, the performance of all three algorithms decreases. However, compared with the other two algorithms, LEGAOCD still achieves higher NMI values, indicating that the proposed algorithm performs well in partitioning synthetic dataset networks.

The ability to detect overlapping community structures does not necessarily align with the ability to detect overlapping nodes [18]. F-score comparison results are shown in [Figure 3: see original paper]. LEGAOCD exhibits high recall and low precision because its post-processing mechanism may detect more overlapping nodes than actually present. Therefore, the algorithm shows average performance in detecting overlapping nodes. Larger O_m values indicate better ability of LEGAOCD to detect overlapping nodes.

All three algorithms achieve SQ values above 0.5 on real-world dataset networks, indicating clear network community structures. As shown in [Figure 4: see original paper], LEGAOCD obtains higher SQ values than other algorithms on real-world dataset networks (except for PGP). In the larger PGP network, LEGAOCD achieves higher SQ values than CPM but lower than COPRA.

4 Conclusion

This paper proposes the LEGAOCD algorithm for overlapping community detection, with main contributions in three aspects: (1) an extended genetic algorithm to solve overlapping community detection problems; (2) a defined motif stability measure to quantify community stability; and (3) a node-community closeness formula serving as the fitness function for the genetic algorithm. For large-scale networks, LEGAOCD and CPM algorithms suffer from excessive time and space overhead. How to mine clear community structures in large networks will be the direction of future research.

References

[1] Zhou Xu. Research on community detection algorithms in complex networks [D]. Changchun: Jilin University, 2016.
[2] Mahmoud H, Masulli F, Rovetta S, et al. Community detection in protein-protein interaction networks using spectral and graph approaches [C]// Computational Intelligence Methods for Bioinformatics, and Biostatistics. [S. l.]: Springer International Publishing, 2013: 62-75.
[3] Palla G, Derényi I, Farkas I. Uncovering the overlapping community structure of complex networks in nature and society [J]. Nature, 2005, 435(7043): 814-8.
[4] Gregory S. Finding overlapping communities in networks by label propagation [J]. New Journal of Physics, 2009, 12(10): 2011-2024.
[5] Xie J, Szymanski B K, Liu X. SLPA: Uncovering Overlapping Communities in Social Networks via a Speaker-Listener Interaction Dynamic Process [J]. 2011: 344-349.
[6] Qiao Shaojie, Guo Jun, Han Nan, et al. Large-scale complex network community parallel discovery algorithm [J]. Chinese Journal of Computers, 2017, 40(3): 687-700.
[7] Han Zhongming, Tan Xusheng, Chen Yan, et al. NCSS: A fast and effective complex network community partitioning algorithm [J]. Scientia Sinica (Informationis), 2016(4).
[8] Prat-Pérez A, Dominguez-Sal D, Brunat J M. Shaping Communities out of Triangles [J]. Computer Science, 2012: 1677-1681.
[9] Zhang Haiyan, Liang Xun, Zhou Xiaoping. A local extension overlapping community detection algorithm for directed graphs [J]. Journal of Data Acqui-

sition and Processing, 2015(3): 683-693.
[10] Qiao Shaojie, Han Nan, Zhang Kaifeng, et al. Overlapping community detection algorithm in complex network big data [J]. Journal of Software, 2017, 28(3): 631-647.
[11] Atay Y, Kodaz H. A New Adaptive Genetic Algorithm for Community Structure Detection [M]// Intelligent and Evolutionary Systems. [S. l.]: Springer International Publishing, 2016.
[12] Wang Qi, Wen Zhiping. An overlapping community detection method based on multi-dimensional genetic algorithm [J]. Computer Application Research, 2016, 33(12): 3543-3546.
[13] Niu Xinzheng, Si Weiyu, She Kun. Dynamic network community detection based on evolutionary clustering [J]. Journal of Software, 2017, 28(7): 1773-1789.
[14] Shen B, Wang N, Qiu H. A new genetic algorithm for overlapping community detection [C]// Proc of the 10th International Conference on Intelligent Information Hiding and Multimedia Signal Processing. Washington DC: IEEE Computer Society, 2014: 766-769.
[15] Li Z, Liu J. A multi-agent genetic algorithm for community detection in complex networks [J]. Physica A: Statistical Mechanics and Its Applications, 2016, 449: 336-347.
[16] Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms [J]. Physical Review E: Statistical, Nonlinear & Soft Matter Physics, 2008, 78(2): 046110.
[17] Zhou X, Liu Y, Zhang J, et al. An ant colony based algorithm for overlapping community detection in complex networks [J]. Physica A: Statistical Mechanics and Its Applications, 2015, 427: 289-301.
[18] Xie J, Kelley S, Szymanski B. Overlapping community detection in networks: the state of the art and comparative study [J]. ACM Computing Surveys, 2013, 45(4): 1-37.

Note: Figure translations are in progress. See original paper for figures.