Complex Network Community Detection by Improved Nondominated Sorting Genetic Algorithm

Wen-jun LIU¹,* and Bin CHEN²

¹School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China
²Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan 430065, China
*Corresponding author

Keywords: Complex networks, Community detection, NSGA2, LPA, Modularity.

Abstract. Aimed at the problems of low solution precision and easy to be trapped into local optima by single objective evolutionary algorithm, a self-adaptive multi-objective optimization algorithm based on nondominated sorting genetic algorithm II (NSGA2) and Label Propagation Algorithm (LPA) is proposed. The algorithm takes Kernel K-means (KKM) and Ratio Cut (RC) as the objective functions. Two new crossover operator and the improved mutation operator is used to achieve the evolution of the population. We conducted simulation experiments in the computer-generated networks and the real-world networks environment. The results show that compared with other community detection algorithms, our algorithm has the advantages of high resolution and strong search ability, and it can effectively identify the community structure in complex networks.

Introduction

Real-world systems are complex, but they can be described by abstract network models, such as ecosystem, social network and so on. By analyzing the community structure in complex network, features in complex networks can be found more intuitively [1,2,3].

Community detection can be modeled as an optimization problem. For example, Pizzuti has proposed a single objective genetic algorithm(GA-net) for community detection[4].Because of the limit of single-objective evolutionary method based on modularity several multi-objective evolutionary methods has been proposed, such as MOGA-net [5].

In this paper, we propose a novel community algorithm based on NSGA2 and LPA, termed as NSGA2-LPA. We use the multi-objective evolutionary algorithm based on NSGA2 which was proposed by Deb in 2002. NSGA2 has high computational efficiency when dealing with low-dimensional problems, and the solution set has good diversity and convergence [6,7].

Related Background

Complex Network

A complex network is usually represented by $G = (V, E)$, where $V$ represents the set of all nodes in the complex network, and $E$ represents the set of all edges in the complex network. The topology of complex networks is usually represented by adjacent matrix $A = (a_{ij})_{n \times n}$, where $n$ is the number of nodes in the network, $a_{ij} = 1$, if there is a link between node $i$ and $j$, otherwise, $a_{ij} = 0$.

Optimization Problem

The community detection algorithm based on improved NSGA2 selects KKM as the evaluation function of the community’s internal density and RC as the evaluation function of the community’s external density [8]. Given an unsigned network denoted as $G = (V, E)$ with $|V| = n$ nodes and $|E| = m$ edges. The adjacency matrix is $A$. The $k$ partitions are in the network denoted...
as $C = \{C_1, C_2, \cdots, C_k\}$, Then KKM and RC can be defined as

$$\min \begin{cases} 
KKM = 2(n - k) - \sum_{i=1}^{k} \frac{L(C_i, C_i)}{|C_i|} \\
RC = \sum_{i=1}^{k} \frac{L(C_i, \bar{C}_i)}{|C_i|} 
\end{cases}$$

(1)

where $L(C_i, C_j)$ is the number of edges between nodes in community $i$; $L(C_i, \bar{C}_i)$ is the number of edges between nodes in community $i$ and nodes outside community $i$.

The Proposed NSGA2-LPA for Community Detection

Encoding and Initialization

We adopt character-based encoding and each gene represents the community number of the node. The character-based encoding process is shown in the Fig. 1. Fig. 1(a) is the network topology, and Fig. 1(b) is the code of nodes in the network. The community number for nodes 1, 2, and 3 are all 1, so they belong to the same community. Fig. 1(c) is a schematic diagram of network community structure, node 1, 2 and 3 form the same community, and node 4, 5, 6, 7 form the same community.

![Figure 1. (a) Network structure diagram. (b) Gene encode table. (c) Community structure map.](image)

This paper adopts the initialization method of randomly assigning community numbers to each node, and randomly generates a population composed of $n$ individuals.

Crossover

(a) Multipoint crossover operator based on community structure

Two chromosomes $A$ and $B$ are randomly selected as the crossed source and target chromosomes respectively which are shown in the Fig. 2(a). Randomly select two nodes from the source chromosome $A$ and get the node numbers marked a and b, get the community structure from a to b on source chromosome $A$ and use the community structure to change the region from a to b on target chromosome $B$ to generate a new chromosome as an individual in offspring population. Exchange the source and target chromosomes to generate another new chromosome.

![Figure 2. Two crossover operator.](image)

(b) Single point crossover operator based on LPA

Also two chromosomes $A$ and $B$ are randomly selected as the crossed source and target chromosomes respectively which are shown in the Fig. 2(b). Randomly select one node from the source chromosome $A$ and get the node numbers marked c, get the community structure from c on source chromosome $A$ and use the community structure to change the corresponding nodes’ numbers on target chromosome $B$ to generate a new chromosome as an individual in offspring population. Exchange the source and target chromosomes to generate another new chromosome.
Mutation

In the process of mutation, we randomly pick a chromosome to be mutated. And a node is picked randomly on the chromosome to be the change-node. Randomly select a point that has a link with the change-node as the mutation guidance node, and change the community number of the change-node to the community number of mutation guidance node.

The Main Loop of the NSGA2-LPA Algorithm

Given a network \( N \) and the graph \( G \) modeling it, NSGA2-LPA optimizes the two objectives (1) presented in Section 2.2. The algorithm starts with a population initialized at random, and then applies the modified crossover and mutation operators to produce the new population. Find the Pareto front set by using the nondominant sort of NSGA2, and the crowding distance is used to select the optimal set as the offspring population. After that, we adopt the concept of modularity, introduced by Girvan and Newman to assess the quality of a partition, to select, among the solutions found, that having the highest value of modularity.

Experimental Results

In order to verify the performance of the algorithm proposed in this paper, Experiments have been carried out on both computer synthetic networks and six real-life networks. And four algorithms named as MOGA-net[9], GA-net[10], IGA[11], GN[1] are chosen to compare with the proposed approach. All experiments were carried out under the desktop PC of Inter(R)Core(TM)i5-4460 processor, 3.2ghz main frequency, 4g memory and Windows 10 operating system.

Synthetic Data Set

LFR(lancichinetti fortunato-radicchi) benchmark network is currently the most commonly used computer-generated data set in community detection which is proposed by Lancichinetti[12]. The LFR network model used in this paper takes as the parameter is:

\[
N = 128, k = 16, k_{max} = 16, c_{max} = 32, c_{min} = 32, \mu \text{ increases from 0 to 0.5 in steps of 0.05.}
\]

The first evaluation index used in this paper is the function modularity \( Q \) proposed by Newman in [1] which is widely used to measure the compactness of community division. The second evaluation index is normalized mutual information (NMI) which is used to evaluate the similarity between the true community structure and the detected ones[13,14].

![Figure 3](image)

Figure 3. (a)Comparing NSGA2-LPA with different algorithms in terms of nmi on LFR benchmark (b)Comparing NSGA2-LPA with different algorithms in terms of Q on real-life network benchmark.

In order ensure the fairness of the algorithm, each algorithm was run independently for 30 times, and the curve of nmi variation with the mixed parameter \( \mu \) obtained by the final statistical average was shown in the figure. As can be seen from Fig. 3(a), when \( \mu > 0.1 \), the algorithm proposed in this paper and GN algorithm can correctly identify the real community structure of the network. The
performance of GN algorithm decreases when $\mu=0.4$, and the performance of our algorithm decreases when $\mu=0.5$. Therefore, the NSGA2-LPA algorithm proposed in this paper has advantages in accuracy and efficiency in LRF network test.

Real-life Data Set

Table 1. Information of real network data sets.

<table>
<thead>
<tr>
<th>Network</th>
<th>Node</th>
<th>Edge</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>karate</td>
<td>34</td>
<td>78</td>
<td>Zachary’s karate club</td>
</tr>
<tr>
<td>dolphin</td>
<td>62</td>
<td>159</td>
<td>Dolphin social network</td>
</tr>
<tr>
<td>book</td>
<td>105</td>
<td>441</td>
<td>Books about US politics</td>
</tr>
<tr>
<td>football</td>
<td>115</td>
<td>613</td>
<td>American College football</td>
</tr>
<tr>
<td>netscience</td>
<td>1461</td>
<td>2742</td>
<td>Coauthorship network</td>
</tr>
<tr>
<td>power grid</td>
<td>4941</td>
<td>6594</td>
<td>American Power Grid</td>
</tr>
</tbody>
</table>

In Fig. 3(b), we show the modularity obtained by each algorithm, and the abscissas 1-6 represent the karate network, the dolphin network, the books network, the football network, the netscience network, and the power grid network [15,16,17,18] which are detailed description is shown in table 1. It can be seen from the experiment that in most cases, the evolutionary algorithm is higher than the modularity solved by the traditional algorithm like GN algorithm, indicating that the evolutionary algorithm is effective for community detection. And in large complex networks, GN algorithms cannot successfully divide the community structure due to the limitations of the algorithm. Compared to single-objective evolutionary algorithms, multi-objective algorithms improve the accuracy of partitions over most networks, especially for large complex networks.

Dolphin dataset is a network of dolphins that Lesseau et al. observed in a group of bottlenose dolphins inhabiting the Doubtful Sound Bay in New Zealand after seven years. A tie between two dolphins was established by their statistically significant frequent association. The network has 62 nodes and 159 edges. The network naturally is separated into two large groups, the female group and the male one as shown in Fig. 4(a). From the Pareto front set of NSGA2-LPA, we can get a high modularity value of the dolphin network, $Q=0.5263$. And the partitioning with highest $Q$ detected by NSGA2-LPA is shown in Fig. 4(b). It is clearly known from this experiment, NSGA2-LPA is a more effective algorithm for community detection.

Concluding Remarks

In this paper, we propose a new community detection algorithm, NSGA2-LPA. This algorithm is inspired by the NSGA2 algorithm, combined with the idea of label propagation, proposes a new crossover operator and mutation operator based on traditional evolutionary operators. It avoids the traditional evolutionary algorithm to destroy the original community structure when searching, and ensures the effectiveness of algorithm search. The algorithm has the characteristics of high
precision and wide applicability. It is more suitable for large complex networks than GN algorithm and it has better accuracy than evolutionary algorithms such as GA-net, IGA, and MOGA-net. It also shows that the proposed method can efficiently reveal community structures in large complex networks.

References


