Applied Soft Computing 19 (2014) 121–133

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Multi-level learning based memetic algorithm for community detection

Lijia Ma, Maoguo Gong*, Jie Liu, Qing Cai, Licheng Jiao

Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University, Xi'an 710071, China

ARTICLE INFO

Article history:
Received 29 March 2013
Received in revised form
16 December 2013
Accepted 2 February 2014
Available online 14 February 2014

Keywords:
Memetic algorithm
Genetic algorithm
Community structure
Complex networks
Modularity

ABSTRACT

Complex network has become an important way to analyze the massive disordered information of complex systems, and its community structure property is indispensable to discover the potential functionality of these systems. The research on uncovering the community structure of networks has attracted great attentions from various fields in recent years. Many community detection approaches have been proposed based on the modularity optimization. Among them, the algorithms which optimize one initial solution to a better one are easy to get into local optima. Moreover, the algorithms which are susceptible to the optimized order are easy to obtain unstable solutions. In addition, the algorithms which simultaneously optimize a population of solutions have high computational complexity, and thus they are difficult to apply to practical problems. To solve the above problems, in this study, we propose a fast memetic algorithm with multi-level learning strategies for community detection by optimizing modularity. The proposed algorithm adopts genetic algorithm to optimize a population of solutions and uses the proposed multi-level learning strategies to accelerate the optimization process. The multi-level learning strategies are devised based on the potential knowledge of the node, community and partition structures of networks, and they work on the network at nodes, communities and network partitions levels, respectively. Extensive experiments on both benchmarks and real-world networks demonstrate that compared with the state-of-the-art community detection algorithms, the proposed algorithm has effective performance on discovering the community structure of networks.

1. Introduction

With the advancement of Internet and Web 2.0 techniques, many real-world complex systems, including online communities, power grids, collaboration systems, disease control systems, resource distribution systems and information recommendation systems, are closely related to our daily activities by sharing information [1–4]. In a world of the huge and disordered information, how to use an effective computing model to mine and analyze the potentially useful information has become a commonly concerning issue [4]. In recent years, the research on complex networks has attracted more and more attentions in the fields of biology, physics, sociology and mathematics [1–3,5]. The topological structure of complex systems can easily be modeled as a complex network with linked nodes. More specifically, the entities of complex systems can be represented as the nodes (or vertices) of networks, and the relations between entities can be modeled as links (or edges) of networks. Community structure is a common and important property in both complex networks and complex systems [6]. In complex networks, communities are made up of a set of nodes which have more connections with each other than those with the remaining nodes in the network [7–10]. The community structure property is indispensable to reflect the potential structural behavior of networks [6]. In complex systems, communities are composed of a few entities which have similar properties. Generally, the potential functionality of complex systems is related to its community structure property [6].

Various methods have been proposed to detect communities in complex networks. Among them, one of the most popular techniques is based on the optimization of an objective, modularity, which is the most widely used criterion to evaluate the quality of the community structure of networks [6,11]. The modularity optimization methods to detect communities are based on searching a particular network partition which has the maximal modularity [6,11]. Recent studies in [12,13] demonstrate that the number of local maxima in the optimization of modularity is exponentially growing with the increase of the size of complex networks.

Many heuristic algorithms based on optimizing modularity for community detection have been proposed in recent years [14–26]. Some of them optimize from one solution to a better one and thus easily get into local optimal solutions [27]. Moreover, some of them are sensitive to the optimized order and thus the detected network

* Corresponding author. Tel.: +86 029 88202661; fax: +86 029 88201023.
E-mail address: gong@ieee.org (M. Gong).

1568-4946/$ – see front matter
http://dx.doi.org/10.1016/j.asoc.2014.02.003

partitions are not stable [6,27]. For instance, the fast greedy method (FM) [14], which iteratively joins a pair of communities with the largest gain in modularity, tends to obtain quite large communities and neglect small ones [6], and thus it is easy to get into a local optimal network partition. The modularity-specific label propagation algorithm (LPAm), proposed by Barber and Clark [19], is sensitive to the optimized order of the nodes, and thus the revealed network partitions are different at different independent trials. Moreover, it is also easy to get into a local optimal network partition where the detected communities are similar in total degree [20]. The multistep techniques, proposed by Blondel et al. [17] and Schuetz and Cafisch [18], which iteratively merges a set of nodes and communities, are also easy to get into a local optimal solution. This is because the merged nodes and communities can hardly be separated again [28]. Several techniques have been proposed to enhance both the accuracy and the stability of modularity-based optimization methods. Lai et al. [29] and Yan and Gregory [30] use network preprocessing techniques to improve the accuracy. Both of them are based on the assumption that the vertices in the same community possess more similar behaviors than those in different communities. Lancichinetti and Fortunato [31] adopt the consensus clustering technique to enhance both the accuracy and the stability of the resulting network partitions. Pizzuti, Shi and we adopt multiobjective optimization algorithms to discover a proper and stable network partition which has a high value of modularity [25,32,33]. In this study, we try to use an intelligent hybrid technique, memetic algorithm (MA), to improve the performance of modularity-based community detection algorithms by simultaneously evolving a population of solutions to better ones.

With the rapid development of computer science, mathematics and biology, the research on intelligent algorithms for solving practical engineering problems has attracted increasing attentions in recent years. Many intelligent algorithms, including genetic algorithm (GA) [34], memetic algorithm [35–37], artificial neural network [38], simulated annealing [15], swarm intelligence algorithms [39], fuzzy system [40–44], artificial immune system [39], and so on, have been designed as problem-specific techniques for tackling and solving real-world applications [32,45–51]. Compared with the traditional algorithms, the problem-specific intelligent algorithms can effectively find a proper solution with high quality in a reasonable period of time [39].

Memetic algorithms (MAs) are hybrid global-local heuristic search methodologies [35]. The global heuristic search is usually a form of population-based method, while the local search is generally considered as an individual learning procedure for accelerating the convergence to an optimum [35]. In general, the population-based global search has the advantage of exploring the promising search space and providing a reliable estimate of the global optimum [36]. However, the population-based global search is difficult to discover an optimal solution around the explored search space in a short time. The local search is usually designed for accelerating the search and finding the best solutions on the explored search space. Therefore, this hybridization, which synthesizes the complementary advantages of the population-based global search and the individual-based local search, can effectively produce better solutions [37]. Recent studies on MAs have demonstrated that they are effective and efficient for tackling the optimization problems in many real world applications [35–37,48–51]. MAs have also been used for uncovering communities in networks in recent years [45,52–54]. For instance, in [45], we proposed a memetic algorithm, named as Meme-Net, to uncover communities at different hierarchical levels. Meme-Net shows its effectiveness. However, its high computational complexity makes it impossible to search communities on slightly large networks. Shang et al. [52] and we [53] try to adopt simulated annealing as an individual learning procedure to decrease the computational complexity of Meme-Net. However, their computational complexity are still very high relative to classical modularity-based community detection algorithms. The algorithm in [54] adopts the technique in [17] as the local search. Meanwhile, it takes a large amount of time and energy on generating initial population as it directly use the algorithm in [17] to initialize a population of solutions. Therefore, the algorithms in [45,52–54] are difficult to apply to real-world problems.

Motivated by the above descriptions, in this study, we present a fast memetic algorithm with multi-level learning strategies to detect communities by optimizing modularity. We term the proposed algorithm as MLCD for short. MLCD adopts a genetic algorithm as the global search and uses the proposed multi-level learning algorithms to accelerate the convergence. The proposed multi-level learning strategies work on the network at node, cluster, and network partition levels, respectively. By iteratively executing GA and multi-level learning strategies, a network partition with high modularity can be accurately and stably obtained. We also employ a modularity-specific label propagation rule to update the cluster identifier of each node at each operation. The simple update rule guarantees the rapidity of the proposed algorithm. Experiments on GN and LFR benchmarks and 12 real-world networks demonstrate that compared with the classical community detection algorithms, MLCD has the superior performance in stably finding a proper community structure of networks. It is also shown that compared with Meme-Net, MLCD takes much less time to find a more proper and stable community division of networks.

The remainder of this paper is organized as follows. In the next section, the problem definition is given. Section 3 gives a detail description for the proposed algorithm. In Section 4, experiments on GN and LFR benchmarks and 12 real-world networks are given to demonstrate the effectiveness of the proposed algorithm. Finally, the conclusion is given.

2. Problem definition

Let us consider an unweighted and undirected graph $G = (V; E)$ which has $|V| = n$ vertices (or nodes) and $|E| = m$ edges (or links), the connection of the graph G can be represented as an adjacency matrix A. Its element A_{ij} is 1 when a link between nodes v_i and v_j exists and 0 otherwise [6]. In order to detect the underlying structure of complex networks, we need to know the definition of community. However, there is no agreement of the definition of community in networks. Its definition is closely relate to the specific systems and practical applications [6]. In most cases, an algorithmic definition of community, which is based on the degree of vertex in graph, is adopted [7]. An algorithmic community (or a sub-graph $S \in G$) is defined as a cluster whose internal degree, the total weight of the links within the sub-graph ($k_S^{in} = \sum_{i\in S, j\in S} A_{ij}$), is larger than its external degree, the total weight of links towards the rest of the network ($k_S^{out} = \sum_{i\in S, j\notin S} A_{ij}$) [7].

Note that, many network divisions satisfy the above algorithmic definition. Most of these divisions cannot reflect the potential structural organizations and functional behaviors of complex systems. In real-world applications, good network divisions can effectively illustrate their potential structural organizations. Therefore, it is necessary to adopt a criterion to evaluate the quality of network partitions.

Many criteria have been proposed to evaluate network partitions, and they are devised according to the knowledge in the fields of physics, biology and sociology. Modularity, proposed by Grivan and Newman [8] based on the graph theory, is up to now the most representative quality criterion for measuring the quality of network partitions. Modularity measures the difference between the actual fraction of edges within communities and its expected value