Community Detection in Social Network Using Artificial Bee Colony with Genetic Operator

Thet Thet Aung
University of Computer Studies, Yangon
thetthetaung@ucsy.edu.mm

Abstract

Community detection (CD) plays an important role in analyzing social network features and helping to find out valuable hidden information. Many research algorithms have been proposed to find the best community in the network. But it has many challenges such as scalability and time complexity. This paper proposes a new algorithm, Artificial Bee Colony Algorithm with Genetic Operator (ABCGO) that combines crossover and mutation operators with Artificial Bee Colony algorithm. This paper takes modularity Q as objective function. Compared with five state-of-art algorithms, results on real world networks reflect the effectiveness of ABCGO.

Keywords- Social Network, Community Detection, Artificial Bee Colony, Modularity

1. Introduction

Social network is a theoretical abstraction, useful in social sciences to study the relationships between individuals, groups, or organizations. The connected items are persons or organizations and the ties are interaction or communication between pairs of actor. Real world social networks are usually found to divide naturally into small communities. Generally, a community in a social network is a set of nodes that are densely connected internally, but loosely connected to the rest of this network. In recent years, community detection in a network has become one of the main topics of fields, such as biology, computer science, physics, and applied mathematics. Great applications of CD are to detect suspicious event in telecommunication networks, link prediction, refactoring the software package, recommendation and containment of virus and warm online, recently for criminal detection in large network.

Networks could be modeled as graphs, where nodes represent the objects and edges represent the interactions among these objects. Communities play special roles in the structure-function relationship, and detecting communities can be a way to identify substructures that may correspond to important functions in the network.

With the development of technology, more and more data can be gotten from all kind of social networks, such as facebook, twitter, others. The big data processing is one challenge in many research fields because the efficiency of previous algorithms cannot be guaranteed with the increasing of the data and the network. So, the effective algorithm for community detection is also needed in Social Network.

Network community detection can be viewed as an optimization problem. Due to their inherent complexity, these problems often cannot be well solved by traditional optimization methods. For this reason, nature inspire algorithms have been adopted as a major tool for dealing with community detection problems. In this work, Artificial Bee Colony with Genetic operator (ABCGO) has been used as an effective optimization technique to solve the community detection problems. Modularity is used to measure the community result because it is one of the popular fitness measures for community structures.

The paper is organized as follow. Related works is presented in section 2. Theory background is in section3 and proposed approach is detailed in section 4. Experimental results are shown in section 5. Finally it is concluded in session 6.

2. Related Works

Community detection is similar to a graph partitioning problem. Most of the graph partition methods are based on optimizing a quality function. Hierarchical clustering techniques depend on similar measure between vertices to form clusters. It is classified under two categories, Agglomerative algorithm and Divisive algorithm. In agglomerative algorithm, starts from vertices as separate communities and ends up with a graph as unique community. Divisive algorithm takes the opposite direction of agglomerative. It starts from a graph as one cluster and ends up with clusters containing similar vertices by removing edges in where the authors use betweenness measure to remove iteratively edges from the network to split it into communities [1]. One community detection algorithms proposed is the Girvan-Newman (GN) algorithm, which introduces a divisive method that iteratively removes the edge with the greatest betweenness value [1].

Maps of random walks (Infomap) proposed by Rosvall and Bergstrom [2]. This algorithm is a flow-based and information theoretic clustering approach. It uses a random walk as a proxy for information flow on a network and minimizes a map equation, which measures the description length of a random walker, over all the network clusters to reveal its community structure. Infomap aims to finding a clustering which generates the most compressed description length of the random walks on the network. Optimization-based methods have been considered as the main category. Optimization method can be divided into two categories; single-objective and multi-objective optimization. Both are proved to be efficient and effective for optimization problem.

Hafez, et al. proposed in artificial bee colony algorithm this employs three types of bees to solve the community detection problem and show how the algorithm performance is directly influenced by the use of different community measures [3].

Gema et.al proposed evolutionary clustering algorithm for community detection using graph based information. In their approach, genetic algorithm with fitness that combines different measure of network topology is used for clustering. Binary encoding was used where binary 1 used to denote the node belong to the community [4]

Discrete Particle Swarm Optimization for community detection problem is proposed by Zhou et.al. Modularity density function have been used for objective function in their approach and particle status updating for discrete PSO have been proposed for the community detection problem [5].

Youcef Belkhiri et.al proposed bee swarm optimization for community detection in complex network. This proposed algorithm takes modularity as objective function and k number of bee to create a search area. The algorithm starts with initial solution called reference solution and the taboo list to avoid cycles during the research process [6].

In this paper, genetic operator based ABC with label based encoding will be used. Modularity will also be used for the objective function of the proposed system.

3. Background

In this section includes the definition of social network, community detection, artificial bee colony algorithm and other definition.

3.1. Social Network

In a social network, G(V,E), where V is a set of nodes and E is the edges between the nodes, a community is a group of nodes with tightly connected edges with each other. The nodes in a community show similar characteristics. For example, in social network, people in a community show similar interest to a trend in a community, for example, buying the same products in online marketing.

3.2. Community Detection

A community (also called cluster or module) in a network is a group of vertices having a high density of edges within them and a lower density of edges between groups [7]. Community Detection is the process through which nodes in networks are clustered based on the connection between them. Nodes in community are densely connected and are sparsely connected to other communities.

The purposes of community detection are to understand the interaction between actors, visualize and navigate large network and forming the basis of other tasks such as data mining. By identifying community structure in network can provide knowledge about how network function and topology affect each other.

Community structure is also named as cover of community. It is a set of communities present in network. It is represented as $C=\{c_1,c_2,c_3,c_4,\ldots,c_j\}$. C is the communities structure and c_1, c_2, c_3, c_j are communities. Figure 1, There are two communities $c_1=\{1,2,3,4\}$ and $c_2=\{5,6,7,8,9\}$.

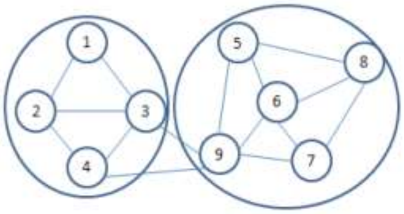


Figure1. Example of communities present in a network

3.3. Artificial Bee Colony Algorithm

Artificial bee colony is one of the most recently defined algorithms by Dervis Karaboga in 2005[8], motivated by the intelligent behavior of honey bees. ABC as an optimization tool provides a population based search procedure in which individuals called foods positions are modified by the artificial bees with time and the bee's aim is to discover the places of food sources with high nectar amount and finally the one with the highest nectar. It contains three groups: scouts, onlookers, and employed bees. Onlookers and employed bees carry out the exploitation process in the search space. Scouts control the exploration process.

The general algorithmic structure of the ABC optimization approach is given as follows:

```
Initialization Phase
REPEAT
    Employed Bees Phase
```

```
        Onlooker Bees Phase
        Scout Bees Phase
        Memorize the best solution achieved so far
    UNTIL (Cycle = Maximum Cycle Number or a Maximum CPU time)
```

In the initialization phase, the population of food sources (solutions) is initialized by artificial scout bees and control parameters are set.

In the employed bees phase, artificial employed bees search for new food sources having more nectar within the neighborhood of the food source in their memory. They find a neighbor food source and then evaluate its fitness. After producing the new food source, its fitness is calculated and a greedy selection is applied between it and its parent. After that, employed bees share their food source information with onlooker bees waiting in the hive by dancing on the dancing area.

In the onlooker bees' phase, artificial onlooker bees probabilistically choose their food sources depending on the information provided by the employed bees. For this purpose, a fitness based selection technique can be used, such as the roulette wheel selection method. After a food source for an onlooker bee is probabilistically chosen, a neighborhood source is determined, and its fitness value is computed. As in the employed bees phase, a greedy selection is applied between two sources.

In the scout bees' phase, employed bees whose solutions cannot be improved through a predetermined number of trials, called "limit", become scouts and their solutions are abandoned. Then, the scouts start to search for new solutions, randomly. Hence, those sources which are initially poor or have been made poor by exploitation are abandoned and negative feedback behavior arises to balance the positive feedback.

These three steps are repeated until a termination criteria is satisfied, for example a maximum cycle number or a maximum CPU time [9]. This paper used the backbone of Artificial Bee Colony and combined with genetic operators.

3.4. Objective Function

Many objective functions for community detection that can capture the intuition of communities have been introduced from different research fields. Quality functions can be used when there is no ground truth for the communities to assess the quality of detected communities. Some of the objective functions such as conductance, expansion, cut ratio, community score and modularity that are already widely used in community detection literatures or can be used for community detection [10]. In this paper, modularity is used to measure the quality of result community.

Newman and Girvan first defined a measure known as 'modularity' to judge the quality of partitions or communities formed. The modularity measure proposed by them has been widely accepted and used by researchers to gauge the goodness of the modules obtained from the community detection algorithms with high modularity corresponding to a better community structure. Network modularity function, also called Q-function, is widely used to quantitatively evaluate the community partition of complex networks.

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \qquad 1$$

In equation 1, where m is the number of edges in network, k_i and k_j are the degree of nodes i and j respectably. A_{ij} is Adjacency Matrix. c_i and c_j denotes the communities of nodes i and j respectably. Function δ (c_i ,c_j) is 1 if i and j are in the same community, otherwise it is 0. The value of Q is between -1 and 1 [11].Objective function plays an important role in the optimization process that leads to good solution. There are many objective function have been proposed to capture the intuition of communities.

4. Proposed Algorithm

ABC algorithm is primarily widely used for real optimization problem. For the CD problem, discrete combinational operators are needed. So ABC with genetic operators is inspired from the genetic algorithm for CD that enables us to use the ABC algorithm for combinational optimization problem.

4.1 Artificial Bee Colony with Genetic Operator

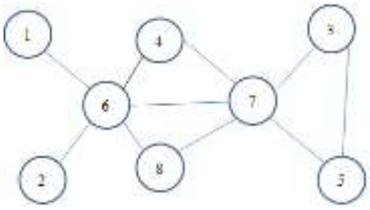
ABC is well known for proven result in numerical optimization problems. Theoretically community detecting problem is NP-hard problem and thus people inclined to choose heuristic algorithms based on objective optimization. Traditional ABC algorithm is suitable for solving real point optimization problem and not suitable for discrete problem like community detection. So, ABC algorithm is modified with genetic operators (crossover and mutation) for community detection. String-based representation is used in the ABC algorithm in which each locus value represents the community index in which it belongs to. Modularity is used as the fitness function of the ABC. In ABC algorithm, exploitation is performed by employed bees and onlookers, while scout bees do the exploration. Employee bee and onlooker bee exploit the food source to create new better food source or solution. In the proposed algorithm, crossover and mutation is used for creating new food source from neighbor.

For the crossover operation, two food source is selected based on their fitness probabilities, one point crossover is performed, from the resulting two food source, greedy selection is applied. Mutation performs the exploration function of the algorithm, in which locus value of food source is randomly mutate to the community index of one of their neighbors allowing the algorithm to explore the search space that have been unexplored.

4.2. Encoded form

According to the nature of community detection problem; a solution is partitioning of nodes V of network G. Each partition contains similar nodes and represents a community. Algorithm starts with initial population creation. An integer array arr is used for data representation of community detection problem. Array store community identifier (CID) of nodes, that arr_i is the community identifier of the node i. The array has n elements and is called as *individual* food source in ABC or *chromosome* in genetic algorithm [12]. There are number of individuals holding different community configuration information in the population. Each individual produces a possible solution.

A possible solution is defined by the number of communities and by the distribution of the nodes in these communities. To represent the assignment of each node in the network for each community, string representation is used. In representation, each locus value indicates the community number the gene belongs to. Figure2 (a) shows a network of 8 nodes to be partitioned: (b) is an initial bee source of this network and the food source of this network represents the reference solution, where node 1, 2,4,6,8 are in community number 1 and nodes 7,3,5 form community number 2.The network is divided into 2 communities and (c) is the community structure of the proposed bees source.

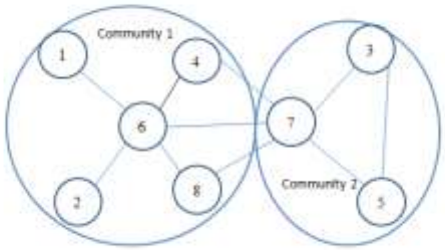


(a)

Food source

Node	1	2	3	4	5	6	7	8
CID	1	1	2	1	2	1	2	1

(b)



(c)

Figure 2: Encoding initial bee strategy for a network

4.3. Genetic Operator

Traditional crossover and mutation of the ABC algorithm need to modify to work with the proposed solution. Crossover uses one point crossover. Pick up a crossover point randomly. Following example shows using graph in figure 2 (a). Choose two individual for crossover.

Food source 1

1	2	3	4	5	6	7	8
1	1	2	1	2	1	1	1

Food Source 2

1	2	3	4	5	6	7	8
1	1	2	2	2	1	2	1

Eg. Crossover point=4

Select gene index with same community as crossover point depend on food source 1. In example, node 4's community id is 1. Then, choose other nodes that community id is the same with node 4.

Food source 1

1	2	3	4	5	6	7	8
1	1	2	1	2	1	1	1

Food Source 2

1	2	3	4	5	6	7	8
1	1	2	2	2	1	2	1

Crossover points [1,2,4,6,7, 8]

Exchange gene values in food source 1 with same gene indexes values in food source 2.

Food Source 1

1	2	3	4	5	6	7	8
1	1	2	2	2	1	2	1

Food Source2

1	2	3	4	5	6	7	8
1	1	2	1	2	1	1	1

Pick up solution with best modularity fitness value. Equation 1 is used in food source 1 and 2 to find better community quality. Food source1's fitness value is 0.195 and food source2's fitness value is 0.12.

For mutation

For each gene, pick up random probabilities, if the rand< mutation probabilities then go on mutation. Example showed, food source 2 gets better fitness value. For genes to be mutated, replace the community index with neighbor node community index. Choose food source 2 to mutate. Node 7 is neighbor with node 4, 6, 8 and 3, the community of node 4,6,8 is 1 and 3 is 2 respectively. It is replaced node 7's community number with 1. This process continues until Maximum Cycle Number is reached.

4.4. Objective function

Objective function plays an important role in the optimization process that leads to good solution. The problem of community detection in Social Network is NP-hard and requires a quality function in order to evaluate and discover good communities. In this work use the modularity Q defined in equation 1. When community group of network is gotten, this community quality is tested by using quality function. This quality function Q assigns a number of each community in a network and communities are ranked based on the score of Q. Good community has higher modularity value. The real network modularity value is generally between 0.3 and 0.7.

5. Experimental Result

In order to show the efficiency of proposed algorithm, choose four real networks used to run. They are Zachary Karate Club [13], Dolphin Network[14] Football Network[15] and Facebook Network [16].

Zachary Karate Club dataset contains friendship network between 34 members of Karate club at US University in 1970. It is one of the most widely used networks in CD. The relationships between members constitute the 78 edges of the network.

Dolphin network is based on the observations of the behavior of 62 dolphins over a period of seven years living in Doubtful Sound, New Zealand. It contains 62 dolphins as nodes and 159 connections as the edges in the network.

American College football game between American colleges during regular season fall 2000. It contains 115 teams as nodes and 613 edges.

Facebook Network dataset in SNAP consists of friends lists from facebook. Facebook data was collected from survey participants using facebok app. It contains 4039 nodes and 88234 edges.

Traditional community detection algorithm is tested on the four real dataset. Some algorithms have good modularity for small network but they can always not get good modularity for large scale graph. Researchers also use population based algorithm to find community detection. This paper uses other five traditional community algorithms [17-21] which are state-of-art method in community detection research.

ABCGO algorithm for community detection is also tested on the four real networks. The results of the algorithm and the other methods are showed in figure 3. Figure 3 illustrates a numeric quality comparison of proposed algorithm with other algorithms, x-axis represents different CD algorithms and y-axis is the modularity value for these datasets. The proposed algorithm gets suitable modularity values in Zachary Karate club, football and facebook datasets. In dolphin, their modularity results are small difference. ABCGO gets excellent result in football dataset because its nodes contain many link connections to the other nodes (densely connected network). The algorithm gets efficient result in large dataset such as football and facebook. The result community quality depends on modularity value. Some algorithms get suitable modularity result but the number of community is different with ground truth. ABCGO uses prior information (the number of community structure) which make the algorithm more targeted and improves accuracy of community detection

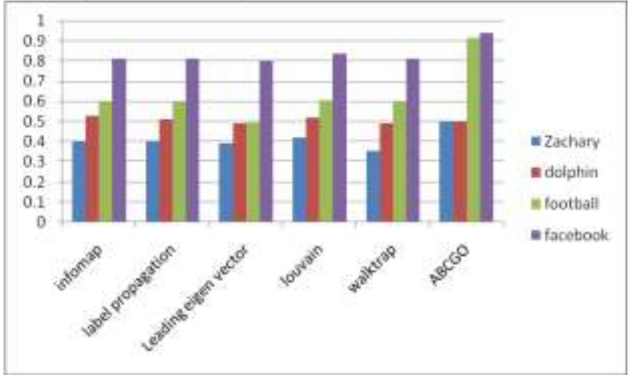


Figure 3. ABCGO on real networks and other algorithms' clustering Modularity quality comparison

.

For proposed algorithm, initial numbers of population and maximum cycle number are important. Population size was 50 and maximum cycle number is 100, crossover rate is 0.6 and mutation rate is 0.1 are used to test the algorithm.

6. Conclusion

Nature inspire algorithms can improved by adjusting balance between exploitation and exploration. This paper proposed a new approach for community detection in social networks called ABCGO. The approach is based on population algorithm to find the community structure and the merging of communities. The approach is constructed for undirected and un-weighted networks. Results obtained on real social network argue for the capacity of the approach to detect real communities. For large scale network dataset, the efficiency and scalability of a community detection algorithm is crucial for its popularity. Future development focuses on the scalability of community detection problem with parallel ABCGO and set some criteria for increasing the accuracy.

7. References

[1]. M. Girvan and M. E. J. Newman.”Community structure in social and biological networks.” Proceedings of the National Academy of Sciences, 99:7821–7826, 2002.

[2]. M. Rosvall and C. T. Bergstrom. "Maps of random walks on complex networks reveal community structure". Proceedings of the National Academy of Sciences of the United States of America, 105(4):1118–23, Jan. 2008.

[3]. Ahmed Ibrahem Hafez, HossamM.Zawbaa, Aboul Ella Hassanien, Aly A. Fahmy, “ Network Community detection using artificial bee colony swarm optimization”. http://scholar.cu.edu.eg/sites/default/files/abo/files/ibica2014_p27.pdf

[4]. Gema Bello-Orgaz , David Camacho, “Evolutionary clustering algorithm for community detection using graph-based information” ,2014 IEEE Congress on Evolutionary Computation (CEC) July 6-11, 2014, Beijing, China

[5]. Zhou, D.-Q ,Wang, X ,Cheng, S.-Y , Chen, Y. (2016). “Community detection algorithm via discrete PSO”. 38. 428-433. 10.3969/j.issn.1001-506X.2016.02.28.

[6]. YoucefBelkhiri, NadjetKamel, HabibaDrias, and SofianeYahiaoui, “ Bee Swarm Optimization For Community Detection in Complex Network”, Spinger International Publishing AG 2017.

[7]. Mehjabin Khatoon, W. Aisha Banu, " A survey on Community detection Methods in Social Networks", I.J. Education and Management Engineering , 2015,1,8-18, May in MECS.

[8]. D. Karaboga, “An idea based on honey bee swarm for numerical optimization”, Technical Report, TR06, ErciyesUniversity,Engineering Faculty, Computer Engineering Department, 2005.

[9]. DervisKaraboga, BeyzaGorkemli, CelalOzturk, NurhanKaraboga, “A comprehensive survey: artificial bee colony (ABC)algorithm and applications”, 11 March 2012.

[10]. Chuan Shi and Yanan Cai, Philip S. Yu, Zhenyu Yan, Bin Wu , "A Comparison of Objective Functions in Network Community Detection", International Conference on Data Mining Workshops © 2010 IEEE

[11]. Mingming Chen, Konstantin Kuzmin, Student Member, IEEE, and Boleslaw K. Szymanski, “Community Detection via Maximization of Modularity and Its Variants”, IEEE trans. Computation Social System, vol. 1(1):46-65, March 2014

[12]. Mursel Tasgin and Haluk Bingol, “Community Detection in Complex Networks using Genetic Algorithm”, 89.75.Fb, 89.20.Ff, 02.60.Gf

[13]. Zachary, W.W, “ An information flow model for conflict and fission in small group.”, J. Anthropol.Res.33,452-473(1977)

[14]. Lusseau, D, “ The emergent properties of a dolphin social network.” Proc.R.Soc.Lond. B Biol. Sci. 270(Suppl 2), S186-S188(2003)

[15]. Girvan,M., Newman, M.E.J, “ Community structure in social and biological networks.”, Proc. Nat. Acad. Sci. 99(12), 7821-7826 (2002)

[16]. “Stanford Large Network Dataset Collection”, http://snap.stanford.edu/data/index.html,

[17]. M. Rosvall and C. T. Bergstrom. "Maps of random walks on complex networks reveal community structure". Proceedings of the National Academy of Sciences of the United States of America, 105(4):1118–23, Jan. 2008.

[18]. Raghavan, U. N., Albert, R. & Kumara, S. “Near linear time algorithm to detect community structures in large-scale networks”.Physical Review E 76, 036106 (2007).

[19]. Newman, M. E. “Finding community structure in networks using the eigenvectors of matrices”. Physical Review E 74, 036104 (2006)

[20]. Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre: “Fast unfolding of communities in large networks”. J. Stat. Mech. (2008) P10008

[21]. Pons, P. & Latapy, M. “Computing communities in large networks using random walks”.In Computer and Information Sciences-ISCIS 2005, 284–293 (Springer, 2005).