Discrete Artificial Bee Colony Algorithm for Community Detection in Social Networks

Thet Thet Aung, Thi Thi Soe Nyunt
University of Computer Studies, Yangon
thetthetaung@ucsy.edu.mm, thithisn@ucsy.edu.mm

Abstract

Community Detection (CD) is an interesting research area in social network analysis. It becomes NP-hard problem in large network. Many researchers use heuristic algorithms to detect communities in social network. These algorithms based on objective optimization function and their goal is to find optimal feasible solution. Artificial bee colony algorithm (ABC) gets high performance and accuracy in continuous problem. To solve discrete problem, ABC should be modified. This paper proposed Discrete Artificial Bee Colony Algorithm (D-ABC) based on the traditional ABC algorithm to solve discrete optimization problem such as community detection. Experiments on three real-world networks show that D-ABC gets more effective community results than other algorithms.

1. Introduction

Social Network (SN) can be modeled as graph in which vertices represent individual (people, group or organization) and links represent the relationship between them. A community is a subgroup of network that have densely connected to each other and loosely connected to the others subgroups. With the development of internet technology, many users connect each other using social network, such as Facebook, Twitter, LinkedIn etc. Determining the communities in social network will be helpful to discover hidden connections between the nodes and, analyzing information flow and understanding the organization and function of the system.

In big data processing, many research fields have challenges to guarantee the efficiency of previous algorithms with the increasing of the data and the network. Numerous algorithms were developed to solve the community detection over the past several years. However, community detection problem is not yet satisfactorily solved. Most of the current algorithms have high computational complexity that makes them unsuitable for large-scale network. So, effective community detection algorithm is needed in social network.

Community detection is discrete optimization problems (combinatorial problem). It becomes a NP-hard problem in large scale network. If the number of nodes increase, whose computational complexity will rise exponentially. The goal of optimization problem is to obtain an optimal solution with respect to a pre-defined objective function. Many researchers proposed community detection algorithms to get more effective, efficient and accurate community result. In the last decades, several optimization algorithms developed based on nature-inspired idea. Nature-inspired algorithms are evolutionary algorithms (Genetic Algorithm, Differential Evolution) and swarm intelligence algorithms (Bat-inspired Algorithm, Ant Colony, Particle Swarm Optimization). Most of them are meta-heuristic based search technology and referred to optimization algorithm. They are applied in wide range of problems.

Artificial Bee colony algorithm is swarm based meta-heuristic algorithm. It is mimic the foraging behavior of honeybee colonies. Karaboga proposed ABC algorithm to solve continuous optimization problem. It gets high performance and accuracy in continuous problem than other heuristic algorithm [1]. When the solution space is discrete type, the traditional ABC should be modified to solve discrete optimization problem. In this paper, discrete artificial bee colony algorithm (D-ABC) is proposed to solve discrete problem especially for community detection problem in social network. It modified initialization and search strategy of bees in ABC to solve discrete problem. Then, replaced fitness function used in ABC by modularity. Modularity is a measure to evaluate the quality of community partitions of network. It is fast and simple in calculation.

The rest of paper is organized as follows: related researches on community detection is shown in section 2. Theory background is described in section 3. In section 4, that provides the proposed approach. The experimental results of D-ABC, that are compared with other algorithms are discussed in section 5. Finally, the paper describes conclusion and future works.

2. Related Works

Many community detection algorithms have been proposed in many research fields over the year. Each has efficient and effective in their way. One popular algorithm is presented by Newman and Girvan which a divide clustering algorithm. In this algorithm, modularity is used as an objective function to judge the quality of resulted community. It iteratively removes edges using their edge-betweenness values until it reaches require modularity value or maximum modularity [2].

Narges Azizifard proposed an algorithm for detecting communities; the goal is to optimize modularity result. Extremal Optimization (EO) method is used, which is divisive algorithm for graph partitioning [3]. Zhou et al. presented discrete PSO for community detection problem and network quality measure function called modularity density is chosen as an objective function. Particle status updating for discrete PSO have also been proposed for detecting community [4]. Saoud Bilal et al. used evolutionary algorithm to find community structure in social network. This algorithm merges community structure to get high modularity value in final result [5].

Many researchers used conventional algorithms for community detection problem [6-13]. Some researchers used bio-inspired swarm intelligence algorithm to detect community. Most algorithms are based on modularity that is foundation measure criterion of community. Community detection becomes a modularity optimization problem. Heuristic algorithms may provide a number of potential solutions to a given problem. Finally, these problem report the best solution to the user. Therefore, many heuristic algorithms have been used for solving the optimization problem.

In this paper, Discrete Artificial Bee Colony Algorithm (D-ABC) with string-based encoding is used to detect community in social network. A community is a subgraph that have dense intra-connection and sparse inter-connection. Algorithms need a measure to find the best community structure that can quantify the quality of community. Modularity measures the quality of divisions of a network into communities. Modularity is also used as the fitness function of D-ABC in this paper.

3. Related Concepts

In this section discusses the related concepts of social network as a graph, artificial bee colony algorithm and quality metrics for measuring the quality of a network partition.

3.1. Graph

Network can be anywhere such as friendship network, follower network, telecommunication network, citation and collaboration network. Graph can be used to represent networks. Social networks are modeled as graph G= (V, E), in which V= {v_1, v_2…, v_n} is set of nodes in the graph and the element of E= {e_1, e_2..., e_m} are links between nodes. In social network, node represents user and link represents the connection between the users. The sample social network is shown in figure 1.

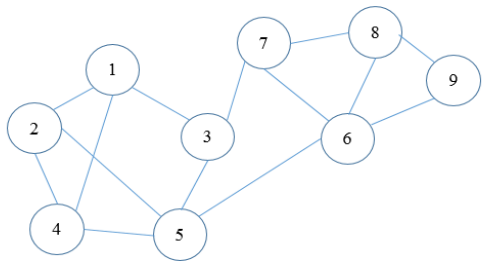


Figure 1: Sample social network

3.2. Community Detection

A community is a group of nodes having a high density of edges between them and lower density of edges between groups. Community detection is the process in which nodes in networks are clustered based on their connectivity and interactively with other nodes. Nodes in the same cluster will have high density of connected edges [14]. The community detection's purposes are to know the relationship between users, visualize network and improve information retrieval. Many community detection algorithms have been developed which differ in term of complexity and network types. Figure 2 shows communities structure present in figure 1 network.

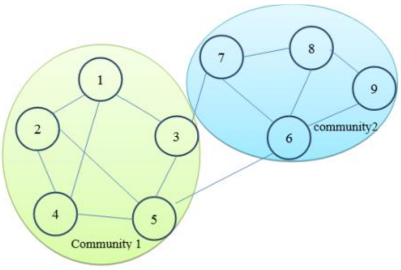


Figure 2. Communities structure present in sample social network

3.3. Quality Metrics

After the network is partitioned into communities, it is needed to judge the quality of result community. Evaluation method can be used if the dataset has no ground truth for communities. It can be used to judge the quality of detected communities [15]. Modularity, inter density, average degree, the fraction over the median degree, conductance, cut ratio, normalize cut etc. can be used to access the quality of communities [16]. Nature-inspired algorithms are based on optimizing the quality function. Therefore, quality function is an important part of solving the optimization problem. In this paper, modularity is used as the fitness function in the proposed algorithm. It also used as evaluation metric to assess the quality of final result communities.

3.3.1. Modularity

Modularity Q is a quality measure metric of network or graph. It quantifies the variance between actual links among all the pairs of nodes within the community and the edges present those pairs in a randomly created model network of the same network. The randomly created network distributes the edges of the actual network randomly among nodes while maintaining the degree distribution of the original network [17]. High modularity means nodes within the community have dense connection but sparse connection to nodes the rest communities. Modularity can be calculated by applying the equation 1.

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

where, m is the number of links in network, k_i is the degree of nodes i and k_j is the degree of nodes j. A_{ij} is 1, if i and j have connection, otherwise it is 0. δ (c_i ,c_j) is a function that is 1 if i and j have same community label, otherwise 0. The value of Q is between -1 and 1. If the modularity value is high, it will get good community structures result.

3.4. Artificial Bee Colony Algorithm

The artificial bee colony (ABC) is a swarm intelligence algorithm that is inspired the foraging behavior of the honeybees. Karaboga introduced it in 2005 for optimizing numerical problems [18]. There are three types of bees; employed bees, onlooker bees and scout bees. Employed bees do for exploiting the nectar sources explored and take the information obtained about the quality of the food source locations that they are exploiting to the onlooker bees.

A food source is chosen by onlooker bee that depending on the information provided by the employed bees. Scouts bee searches new food source depending on an internal motivation or based on possible external traces [19]. After initialize population, food source position is randomly selected by bee and its nectar amount is determined. In ABC, the solutions in the population are equal to the number of the employed bees or the onlooker bees. For each food source, there are only one employed bee. Robust search algorithm requires appropriate balance between exploitation and exploration process. Exploitation process make to refine and reuse existing solution to search better solution that is effective in searching large search space. Exploration widen search to new area. In ABC, employed bees and onlooker bees carried out exploitation process and scout bees make exploration process.

4. The Proposed Algorithm

Karaboga proposed ABC to solve numerical function optimization. In continuous problem, ABC get high performance and accuracy [20]. If the solution space of problem is discrete, ABC needs to modify for solving this problem [21]. In this paper, discrete ABC algorithm is proposed based on traditional ABC to solve combinational optimization problem.

D-ABC employs the network modularity is used the fitness function. In this optimization problem, the number of food sources represents the population of possible solutions. The position of food source indicates the position of a better solution to the given optimization problem. Figure 3 shows the community detection algorithm using D-ABC.

Input: A network G= (V,E), MCN, Pz, Limit
MCN: maximum cycle number
Pz: population size, Limit: trial count
Output: Community membership assignments for network nodes $C=\{C_1,C_2\ldots,C_k\}$
k is the number of communities.
Algorithm:
1.Read input network and initialize the position of food source $X=\{X_1, X_2,X_3\ldots,X_p\}$ using string-based encoding schema.
1.1 For each solution X_i, compute fitness of food source using Q in equation 1.
1.2 Keep the best food source (BestFS) with highest fitness in the population
2. Cycle=1
3. while(Cycle<MCN) {
3.1 Employed Bees Phase
For each employee bee
3.1.1 Generate new solution V_i by applying the neighborhood of X_i
3.1.2 Calculate the fitness of new solution V_i using Q
3.1.3 If (fitness of V_i> fitness of X_i) then replace old solution with new one and reset trial limit of new solution; else increase trial limit count of old solution by 1
3.2 Compute the probability P_i for each X_i using equation 2
3.3 Onlooker Bees Phase:
For each onlooker bee
3.3.1 Generate new solution V_i by applying neighborhood of X_i, depend on the probability of P_i of food source
3.3.2 Calculate the fitness of new solution V_i using Q
3.3.3 If (fitness of V_i>fitness of X_i) then replace old solution with new one and reset trial limit of new solution; else increase trial limit count of old solution by 1
3.4 Scout Bee Phase
3.4.1 If (trial count of X_i> limit) then reset the limit and generate a new solution; else continue
3.5 If (fitness of BestFS < fitness of X_i) then assign the best solution food source X_i as the global best; else continue
3.6 Cycle++} //end while
4. Output the global best solution that contain the community membership assignments of network's node

Figure 3. Community detection algorithm based on D-ABC

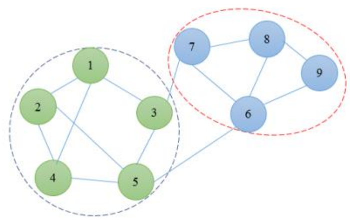
The following parts discuss the encoding schema for solution representation in the population, how to initialize the population of solutions, fitness function for proposed algorithm, how to search the initial population of individual to optimize the quality of communities in individual and finally show how to choose the output in the solutions.

4.1. Solution Representation

Community detection problems mostly use two famous encoding schemas such as locus-based and string-based [22]. In two schemas, a vector of V positions represents individual food source, where V is the number of nodes in the network. The proposed algorithm uses string-based encoding schema. In this representation, each position in the vector is an integer value in the range {1,2,3…, k} where k is the number of groups in the network. Each j^{th} value of food source specifies the cluster label (community identifier CID) they belonged. Figure 4 illustrates the example of string-based encoding schema for a food source. Figure 4 (a) shows the solution representation of individual food source where nodes 1,2,3,4 and 5 belong to community label 0: nodes 6, 7, 8 and 9 form community label 1; (b) is the communities' structure result of the proposed food source.

NodeID	1	2	3	4	5	6	7	8	9
CID	0	0	0	0	0	1	1	1	1

(a) Food source representation



(b) Communities structure of proposed food source

Figure 4. String-based encoding for a solution

4.2. Initialize Population

The population initialization plays an important role. The numbers of food sources represent the number of solutions. The algorithm will accelerate the convergence speed while searching in the search space, if initial population generates better solutions. The proposed algorithm strategy is described as follows:

1. Initialize population of solutions $X=\{X_1,X_2,X_3\ldots,X_p\}$, where p is the population size.
2. For each solution $X_i=\{x_{i1},x_{i2},x_{i3}\ldots,x_{in}\}$, in this vector n represent the number of node in the given network. Every node assigns with same community label.
3. For each solution, a random node is selected. Then assign same community label to random

node and its neighbor nodes. If the randomly choose node does not have neighbors, select another one.

In figure 1 network, population of solutions is {110110000}, {000001011}, {110110000}, {001001110}, {001001110} and {000001111}. In food source {000001111}, nodes 1,2,3,4,5 contain in community label 0 and nodes 6,7,8,9 contain in community label 1. It got two communities. After initialization, evaluate the fitness of each food source.

4.3. Fitness Function

One of the quality metrics, modularity (equation 1) is used as the fitness function in ABC. The objective quality function plays the role of biological fitness in nature-inspired algorithm. It is used to measure the result community quality of individual. Strongly connected communities have higher modularity value. In example, fitness values of {000001111} is 0.346938 and {110110000} is 0.2831632.

4.4. Search Strategy

Initialize population of solutions is evolved by using D-ABC to optimize the quality of communities in each solution. In each cycle, these evaluations have an effective impact on the final optimal result. Community detection is considered as a discrete optimization problem.

When employed bee creates new food source, choose neighborhood food source in the population. Firstly, the genes indexes are randomly choosing in the current food source. Then swap the indexes values of current food source with values in the same indexes of neighborhood food source. Example of creating a new food source is described as follow.

Current food source

1	2	3	4	5	6	7	8	9
1	0	0	0	0	**0**	1	**1**	1

Neighbor food source

1	2	3	4	5	6	7	8	9
0	0	0	0	0	**1**	0	**1**	1

Candidate food source

1	2	3	4	5	6	7	8	9
0	0	0	0	0	1	1	1	1

After creating candidate food source, evaluate the fitness of candidate solution. Choose the best solution between current and candidate solution based on the modularity result. Then calculate each food source probability in population.

Employed bees provide nectar information about the food sources to onlooker bees. An onlooker bee chooses food source depend on its probability value. Probability of each food source can be calculated by using the equation 2. In which, $fitness_i$ is fitness value of X_i and maxFitness is the maximum fitness value of best solution (BestFS).

$$P_i = (0.9*(fitness_i / maxFitness)) + 0.1 \quad \textbf{(2)}$$

After choosing a food source, onlooker bee selects neighborhood food source in the population create new food source as employed bee.

If the solution (food source) is greater than trial limit, the solution is abandoned and scout bee finds new random solution to get global optimal solution. At each cycle, it memorizes the best solution found in the population. Iteration is repeated until the maximum cycle number. The proposed algorithm chooses the best member of food source in the population. Finally, produce the communities of that member as an output. For figure1 network, the proposed algorithm produces two communities (1,2,3,4,5) and (6,7,8,9) which have highest fitness value.

5. Experimental Results

Experiments aim to achieve higher result of modularity. Apply D-ABC on three popular social networks, they are most widely used networks in community detection such as Zachary Karate Club Network [23], Bottlenose Dolphin Network [24], and American College Football network [25]. Zachary Karate network is a network of relationships among the members in karate club. It contains 34 nodes and 78 edges. Bottlenose Dolphin network is a network of 62 dolphins living in Doubtful Sound, New Zealand. It contains 159 links between dolphins. Football network represents schedule of football games between American colleges. Nodes represent teams in football network and links represent match between two teams. It contains 115 teams/nodes and 613 games/edges.

For the proposed algorithm sets population size according to the size of data sets, maximum generation is set 100, limit is set 30. The size of population for karate club network is 50, for dolphin network is 100 and for football network was set to

100. To compare quality result of communities' structure in network, modularity value is also used as evaluation method. The results of the approach are compared with some traditional community detection algorithms label propagation (LP) [8] and walktrap [9] and other heuristic search algorithm such as Genetic Algorithm (GA), Bi-Objective Community Detection (BOCD) [25], Group Search Optimization (GSO) [26], and Evolutionary algorithm (EA) [5].

Table 1 compares the modularity values of community detection algorithms in three social networks. In table 1, D-ABC got effective results on three datasets. Figure 5 presents the modularity results comparison figure based on table1. X-axis shows the name of community detection algorithms and Y-axis is the modularity values of each dataset. The experimental results show that the proposed algorithm effectively detects communities in social networks as other CD algorithms.

Nowadays, nature-inspired algorithms are widely used in CD problem. These algorithms use fitness function which is obtained from quality functions. They also provide a number of potential solutions to a given problem and can choice best solution. Therefore, researchers use population based method to solve community detection problem. The proposed algorithm only considers to get the higher quality modularity result in a given network. It does not consider other factors.

Table 1: Comparison of modularity values in seven community detection algorithms

Algorithm	Modularity Value of Datasets		
	Zachary	Dolphin	Football
LP	0.402	0.511	0.59
Walktrap	0.353	0.48	0.60
GA	0.38	0.495	0.577
GSO	0.384	0.429	0.427
EA	0.38	0.46	0.56
BOCD	0.419	0.507	0.577
D-ABC	0.419	0.52	0.602

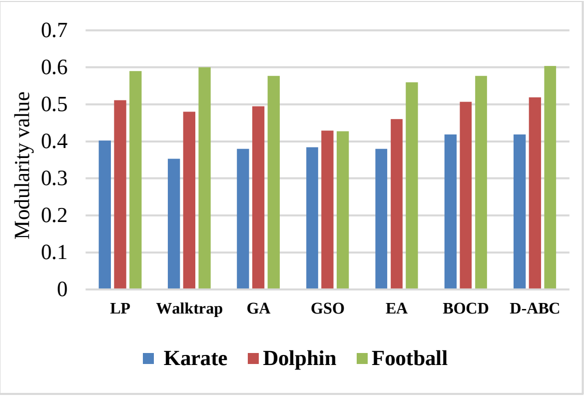


Figure 5: Comparison of modularity results on three social networks

6. Conclusion and Future Work

In this paper, discrete artificial bee colony algorithm has been proposed to solve discrete problem as community detection in social network. Modularity Q is used as fitness function in ABC. String-based encoding schema is applied to represent a community structure. Traditional ABC provided acceptable solution in many numerical optimization problems. The proposed Discrete Artificial Bee Colony Algorithm can also effectively solve the community detection problem in social networks according to the above experimental results. Nowadays, scalable and efficient algorithms are needed to detect community in large-scale social network.

Future work focuses on scalable community detection on large-scale network with parallel D-ABC on spark framework and will be consider other important factors. The real large-scale citation network dataset will also be used to test the effective of the proposed algorithm.

References

[1] Twinkle Gupta, Dharmender Kumar, "Optimization of Clustering Problem Using Population Based Artificial Bee Colony Algorithm: A Review", 2014, IJARCSSE, Volume4, Issue 4, April 2014.

[2] M.Girvan and M.E. J.Newman, "Community structure in social and biological networks." PNAS, 99(12):7821-7826, June 2002.

[3] Narges Azizifard, Mehregan Mahdavi, Babak Nasersharif, "Clustering Social Networks Based on Modularity Optimization", IJIPM, volume3, Number 2, April 2012

[4] Zhou, D.-Q, Wang, X, Cheng, S.-Y, Chen, Y. "Community Detection Algorithm via Discrete PSO", 2016.

[5] Saoud Bilal, Moussaoui Abdelouahab, "

Evolutionary Algorithm and Modularity for Detecting Communities in Networks".
[6] Newman M. "Fast algorithm for detecting community structure in networks." Phys. Rev. E. 69,066133.
[7] Martin Rosvall, Carl Bergstrom. "Maps of random walks on complex networks reveal community structure". Proceedings of the National Academy of Sciences.
[8] Raghavan U.N.; Albert R.; Kumara S "Near linear time algorithm to detect community structures in large-scale networks." Phys. Rev. E. 76, 036106, 2007
[9] Pons, P. & Latapy, M. "Computing communities in large networks using random walks". ISCIS 2005, 284–293 (Springer, 2005).
[10] Bruce Hendrickson and Robertleleand. "A Multilevel Algorithm for Partitioning Graphs". Sandia National Laboratories. "
[11] Holland J.H. "Adaptation in Natural and Artificial Systems." University of Michigan Press, Ann Arbor, Michigan. 1975
[12] Jorg Reichardt and Stefan Bornholdt. "Statistical Mechanics of Community Detection."
[13] Pascal Pons and MatthieuLatapy."Computing Communities in Large Networks Using Random Walks." LIAFA - CNRS and Universit´e Paris 7
[14] Mehjabin Khatoon, W. Aisha Banu, " A survey on community detection method in social network", I.J Education and Management Engineering, 2015,1,8-18, May in MECS.
[15] Josiane Mothe,Karen Mkhitaryan, Mariam Haroutunian, "Community Detection: Comparison of State of the Art Algorithms", 2017,IEEE
[16] Suman Saha and Satya P. Ghrera, "Network Community Detection on Metric Space",26 Aug 2015, www.mdpi.com/journal/algorithms
[17] Nidhi Arora and Hema Banati, "GSO Based Heuristics for Identification of Communities and Their Leaders", © Springer International Publishing AG 2017, DOI 10.1007/978-3-319-65139-2_5
[18] Dervis Karabora, "An idea based on honey bee swarm for numerical optimization". Technical report-tr06, Erciyes University, Engineering Faculty, Computer Engineering Department, Kayseri, Turkiye,2005
[19] B. H. Adebiyi, M. B. Mu'azu, A. M. S. Tekanyi, A. T. Salawudeen, R. F. Adebiyi, "Knowledge-Based Artificial Bee Colony Algorithm for Optimization Problems", Journal of Engineering Research, Volume 22 No 2 December 2017
[20] Akay B, "Performance analysis of artificial bee colony algorithm on numerical optimization problems.", PhD thesis in Turkish, Erciyes University, Graduate School of Natural and Applied Sciences, Kayseri, pp 70–72
[21] Mustafa Servet Kiran, Mesut Gunduz, HazimI, "The analysis of discrete artificial bee colony algorithm with neighbourhood operator on travelling sale man problem". Neural Computing & Application (2013) 23:9–21,
[22] Mohamed Guendouz, "A Discrete Black Hole Optimization Algorithm for Efficient Community Detection in Social Networks", Handbook of Research on Biomimicry in Information Retrieval and Knowledge Management, © 2018, DOI: 10.4018/978-1-5225-3004-6.ch009
[23] Zachary, W.W, "An information flow model for conflict and fission in small group.", J. Anthropol.Res.33,452-473(1977)
[24] Lusseau, D, "The emergent properties of a dolphin social network.", Proc.R.Soc.Lond. B Biol. Sci. 270(Suppl 2), S186-S188(2003)
[25] R.Agrawal, "Bi-objective community detection (BOCD) in network using genetic algorithm," In Contemporary Computing, pp.5-15, Springer Berlin Heidelberg, 2011
[26] S.He, Q. H. Wu, and J.R. Saunder, "Group search optimizer an optimization algorithm inspired by animal searching behavior.", Evolutionary Computation, IEEE transaction on, 13(5), 973 -990,2009