The Eurasia Proceedings of
Educational & Social Sciences (EPESS)
ISSN: 2587-1730

The Eurasia Proceedings of Educational & Social Sciences (EPESS), 2017

Volume 7, Pages 221-226

ICRES 2017: International Conference on Research in Education and Science

DETECTION OF COHESIVE SUBGROUPS IN SOCIAL NETWORKS USING INVASIVE WEED OPTIMIZATION ALGORITHM

Yilmaz Atay
Department of Computer Engineering, Selçuk University

Ismail Koc
Department of Computer Engineering, Selçuk University

Mehmet Beskirli
Department of Computer Engineering, Selçuk University

Abstract: Social network analysis (*SNA*) is a very popular research area that helps to analyze social structures through graph theory. Objects in social structures are represented by nodes and are modeled according to the relations (edges) they establish with each other. The determination of community structures on social networks is very important in terms of computer science. In this study, the Invasive Weed Optimization (*IWO*) algorithm is proposed for the detection of meaningful communities from social networks. This algorithm is proposed for the first time in community detection (*CD*). In addition, since the algorithm works in continuous space, it is made suitable for solving the *CD* problems by being discretized. The experimental studies are conducted on human-social networks such as Dutch College, Highland Tribes, Jazz Musicians and Physicians. The results obtained from experimental results are compared and analyzed in detail with the results of the Bat Algorithm and Gravitational Search Algorithm. The comparative results indicate that *IWO* algorithm is an alternative technique in solving *CD* problem in terms of solution quality.

Keywords: Community detection, discretization, invasive weed optimization, social networks, *SNA*

Introduction

Community structures in network analysis are focused on relationships among people in real-world networks. Detection of community structures in social networks is a problem of considerable interest that has received a great deal of attention (Girvan & Newman, 2002; Leskovec, Lang, & Mahoney, 2010; Papadopoulos, Kompatsiaris, Vakali, & Spyridonos, 2012; Steinhaeuser & Chawla, 2008). Social networks are based on relationships between people, such as political, formal-informal, regional, religious, friendship, business partnership, common social activities, and relational closeness on the internet. The specific properties of social networks need to be well known in order to be able to determine the structural and functional characteristics of network relationships. People in social networks and their relations are represented with vertices and edges, respectively. Graphs which are used to represent any given networks are referred to as the simplest form of undirected networks. Social network analysis (*SNA*) is an approach that analyzes interactions between interrelated social relationships (Makagon, McCowan, & Mench, 2012). *SNA* is also an analysis that examines social structures on network and graph theory. Many different approaches are proposed in the analysis of social networks. Relations of networks of various persons or groups in the face of important events are mathematically formulated and analyzed by *SNA*. In this study, the *CD* approach was examined to reveal and analyze relational information in networks. According to certain features of social networks, the network can be divided into different subnets. For this purpose, several methods are proposed which determine the community structures in the network. Some of these methods refer to certain features in the network and provide various approaches using graph theory. The detection of communities in networks is a field that is long studied and has still maintained its popularity. Contrary to classical methods (Girvan & Newman, 2002; Newman, 2004; Pons & Latapy, 2005), various optimization methods which do not offer a deterministic approach has been applied to the

- Selection and peer-review under responsibility of the Organizing Committee of the conference

*Corresponding author: Yilmaz Atay- Email: yilmazatay@selcuk.edu.tr

CD (Attea, Hariz, & Abdulhalim, 2016; Li & Song, 2013; Shi, Liu, & Liang, 2009; X. Zhou, Liu, Zhang, Liu, & Zhang, 2015). The modularity introduced in the Newman-Girvan algorithm to determine the stopping criterion initially, quickly became an objective function for many optimization methods to determine the communities of complex networks (Li & Song, 2013). After using this criterion as an objective function, many objective functions similar to this function have been proposed (Leskovec et al., 2010). Many studies have been carried out to identify community structures in social networks. The main purpose of our study is to divide the network into divided relational subnets by an objective function. For this purpose, we focused on the invasive weed optimization algorithm (*IWO*). Also, two additional algorithms were used to evaluate the results of *IWO*. These are bat algorithm (*BA*) and gravitational search algorithm (*GSA*). Four different human social networks were used as test datasets.

Background and Motivation

Structurally or functionally meaningful subnetwork structures, referred to as community structures (Newman, 2004) can be determined by optimization of various objective functions. The optimization of objective functions is related to graph clustering according to the *modularity* measure. When any given network presented by graph structure, obtained communities can be considered as subgraphs which have features like maximum common feature in itself, a number of interactions, positional similarities, and such that. Vertices which are the elements of these structures should have maximum interaction and common properties with vertices in their communities and fewer interactions and common properties with vertices in other communities. Groups of people who have a strong relationship in a social environment, the connections of people in a terrorist attack and collaborations of people using computers that make the most data exchanges are some of the examples that can be given about *CD* in *SNA*.

Modularity is the most used objective function in the *CD* problem. Let's assume that G(V, E) represents an undirected and unweighted social network. V and E represent the set of vertices and the set of edges, respectively.

$$V = \{v_i \mid i = 1, 2, 3, \ldots, n\} \text{ and } E = \{e_j \mid j = 1, 2, 3, \ldots, m\} \tag{1}$$

$$Adj_{(i,j)} = \begin{cases} 1 & \text{if node } i \text{ and node } j \text{ are connected,} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

$$Q = \frac{1}{2 \times m} \sum_{ij} \left(Adj_{(i,j)} - \frac{k_i \times k_j}{2 \times m} \right) \times \delta(C_i, C_j) \tag{3}$$

$$m = \frac{1}{2} \sum_{ij} Adj_{(i,j)} \tag{4}$$

$$k_i = \sum_j Adj_{(i,j)} \tag{5}$$

$$\delta = \begin{cases} 1 & \text{if } C_i = C_j \\ 0 & \text{if } C_i \neq C_j \end{cases} \tag{6}$$

i, j, n and, m represent the node index, edge index, number of nodes and, total edges number, respectively. Let's define adjacency matrix as *Adj* with nxn size. This matrix shows the relationships of the elements in the set V by the elements of the set E. *Adj* adjacency matrix is generated by Eq. 2 (Clauset, Newman, & Moore, 2004). Q is named as modularity and expresses the objective function to be maximized. Modularity values are calculated by Eq. 3. Also, m in the Eq. 4 demonstrates the total number of edges in network. k_i demonstrates the degree of i^{th} node, k_j expresses the degree of j^{th} node and as an example k_i can be calculated by Eq. 5. C_i and C_j demonstrate communities of the nodes i and j. δ is a value generated by the function in Eq. 6 which indicates whether nodes i and j are in the same community.

In this work, the modularity Q function was used as fitness function for the detection of the most suitable community structures from the social real-world networks used in the experiments.

Invasive Weed Optimization Algorithm

The invasive weed optimization (*IWO*) algorithm is an metaheuristic optimization technique which has been first introduced by Mehrabian in 2006 (Mehrabian & Lucas, 2006). This algorithm is inspired by the survival of the weeds in the nature by constantly strengthening the lineage. In weed ecology, there is a mechanism in which powerful individuals (weeds) can always survive. These properties of invasive weeds are the basis of the *YOA*. The process of the algorithm can be expressed in 4 sections (Y. Zhou, Chen, & Zhou, 2014):

Initialization of Population

The initial population is created by randomly distributing a certain number of individuals to the problem space.

Reproduction

Each individual needs to derive a certain number of new individuals in proportion to its health condition. The formula of weeds producing seeds is given in Eq. 7.

$$weed\mathrm{i} = \frac{fcurrent - fmin}{fmax - fmin} * (Smax - Smin) + Smin \tag{7}$$

where *fcurrent* is the current fitness of weed. *fmax* and *fmin* express the maximum and the minimum fitness values of the population at the current iteration, respectively. *Smax* and *Smin* respectively identify the maximum and the minimum number of weed value to be produced.

Spatial Dispersal

The determined number of individuals in the previous step is distributed to the problem space as randomized according to Eq. 8 and depending on the time-varying parameter.

$$\sigma_{iter} = \frac{(iter_{\max} - iter)^n}{(iter_{\max})^n}(\sigma_{initial} - \sigma_{final}) + \sigma_{final} \tag{8}$$

where $iter_{max}$ is the maximum number of iterations, σ_{iter} is the standard deviation at the current iteration and n is the nonlinear modulation index.

Competitive Exclusion

Since the number of new generated weeds in the populations will exceed the maximum number of populations after a few iterations, the worst individuals will be eliminated and the number of weeds will be reduced to the maximum number of populations initially determined. In this way, weaker weeds will be eliminated from the population and the stronger weeds will survive. This process continues until all iterations have been completed or until the other stopping criterion reaches the limit value. As a results of these operations, the best weed that has managed to survives gives the best solution for the given problem.

Experimental Results

In this study, three different metaheuristic algorithms have been used in order to make analysis of their performances and compare each other. These algorithms are Invasive Weed Optimization Algorithm (*IWO*), Bat Algorithm (*BA*) and Gravitational Search Algorithm (*GSA*). These algorithms have been examined on four different social networks. These networks are called Dutch College, Highland Tribes, Jazz Musicians and Physicians. All of the experimental studies have been conducted under equal conditions and number of maximum function evaluation (maxFEs) has been taken as 10000. The other parameters of experimental studies have been given in Table 1.

Table 1. Parameter values of experimental studies

Parameter Name	**Value**
Number of Iterations	500
Number of Population	20
Number of Runs	10
Number of FEs	10000
***IWO* Parameters**	
Number of Maximum Seeds	5
Number of Minimum Seeds	1
Number of Maximum Population	40
Variance Reduction Coefficient	2
Initial Value of Standard Deviation	0.5
Final Value of Standard Deviation	0.001

The properties of the social networks used in the experiments have been given in Table 2. As shown in the table, four different social networks which have different number of nodes and edges have been used in the experimental results.

Table 2. The properties of the social networks used in the experiments

Network	**Number of Nodes**	**Number of Edges**
Dutch College	32	3062
Highland Tribes	16	58
Jazz Musicians	198	2742
Physicians	241	1098

The comparative results of *IWO* algorithm with *BA* and *GSA* in terms of solution quality have been presented in Table 3. Since the *CD* problem is the maximization problem, the maximum values are marked as bold. According to this table, it can be stated that *IWO* algorithm is successful than both of *BA* and *GSA* in terms of solution quality for all of the datasets. In addition, it is obviously seen that *IWO* method is quite robust than the other two algorithms. While *IWO* yields results close to results of *BA* and *GSA* only for *Physicians* network, for other networks *IWO* has outperformed *BA* and *GSA* in terms of convergence to the optimal solution. Especially, for *Dutch College* network, *IWO* has clearly outperformed the other algorithms.

Table 3. Comparison of *IWO* algorithm with *BA* and *GSA* in terms of solution quality

ataset/Algorithm	***IWO***		***BA***		***GSA***	
	mean	*std*	*mean*	*std*	*mean*	*std*
Dutch College	**0.0254**	**0.0022**	0.0109	0.0045	0.0090	0.0044
Highland Tribes	**0.1641**	**0.0055**	0.1370	0.0212	0.1566	0.0090
Jazz Musicians	**0.4087**	**0.0090**	0.3458	0.0349	0.3755	0.0180
hysicians	**0.6637**	**0.0014**	0.6547	0.0088	0.6561	0.0059

In terms of processing time, comparative results of *IWO* algorithm with *BA* and *GSA* have been given in Table 4. According to this table, it can be expressed that performance of *IWO* has been superior to those of *BA* and *GSA* since *IWO* has completed each experimental study in shorter time. In addition to this, *IWO* has been more robust among the other algorithms.

Table 4. Comparison of *IWO* algorithm with *BA* and *GSA* in terms of processing time

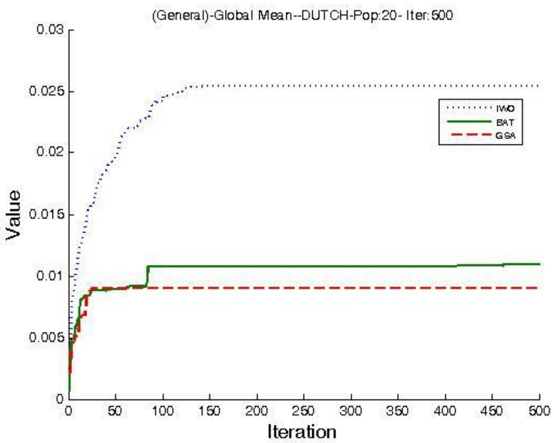
Dataset/Algorithm	***IWO***		***BA***		***GSA***	
	mean	*std*	*mean*	*std*	*mean*	*std*
Dutch College	**39.63**	**0.47**	52.90	1.92	61.83	2.38
Highland Tribes	**10.53**	**0.28**	15.36	0.29	13.26	0.61
Jazz Musicians	**1093.54**	**16.69**	1451.05	77.02	1478.24	22.64
Physicians	**1342.17**	**16.88**	1787.24	41.30	2126.50	39.75

In terms of number of community, comparative results of *IWO* algorithm with *BA* and *GSA* have been reported in Table 5. Minimum values have been marked as bold because it is desirable to have a low number of communities. Considering the results in the table, in terms of number of community, the algorithms have no advantage over each other. According to these results, it can be deduced that the low number of communities may not maximize the modularity value.

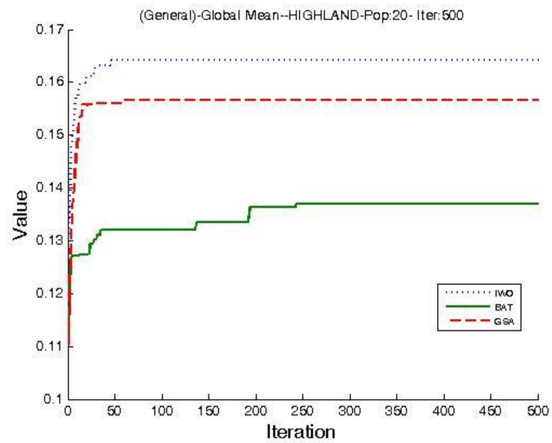
Table 5. Comparison of *IWO* algorithm with *BA* and *GSA* in terms of number of community

Dataset/Algorithm	*IWO*		*BA*		*GSA*	
	mean	*std*	*mean*	*std*	*mean*	*std*
Dutch College	**1.3**	0.483	2.1	0.316	2.0	0.000
Highland Tribes	2.3	0.483	**2.0**	0.000	2.5	0.527
Jazz Musicians	6.2	2.150	4.7	1.636	**4.5**	0.527
Physicians	14.7	2.263	**5.7**	1.059	5.9	1.101

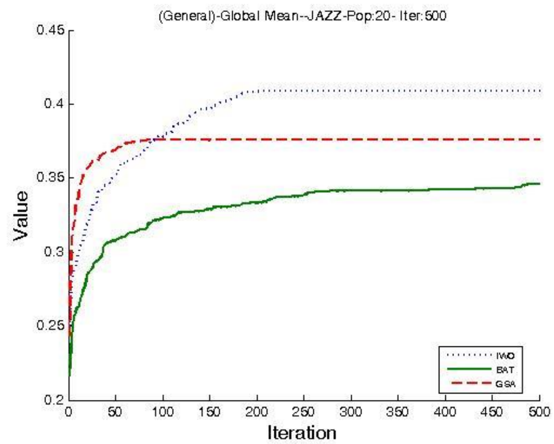
In addition to these experimental results, the convergence curves of the networks have been depicted in Fig. 1. Each experimental study has been performed 10 times for all of the algorithms and the average results have been graphically shown. MaxFEs has been also taken as 10000.



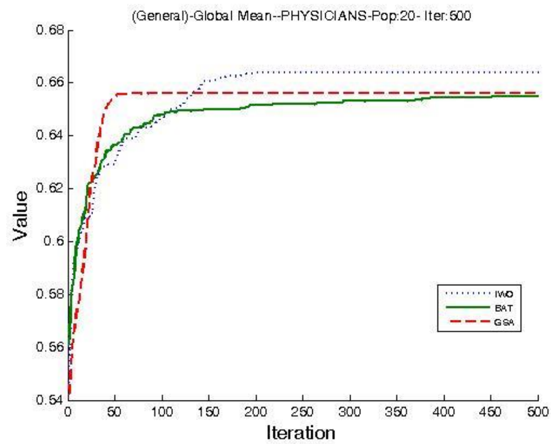
a) *Dutch College* network



b) *Highland Tribes* network



c) *Jazz Musicians* network



d) *Physician* network

Figure 1. Convergence curves for *the* networks

For *Dutch College* network, although all of the algorithms have provided improvement in the first 100 iterations, they could not improve in subsequent iterations. Nevertheless, *IWO* has been clearly superior to other algorithms for convergence speed. For *Highland Tribes* network, although the performance of *GSA* has approached to that of *IWO*, *IWO* has again obtained better results than *GSA*. For *Jazz Musicians* network, although the performance of *GSA* has been better than that of *IWO* in first 100 iterations, *IWO* has outperformed to *GSA* in the end. For *Physician* network, although the performances of *GSA* and *BA* have been better than that of *IWO* in first 50 iterations, *IWO* has outperformed to both of *GSA* and *BA*. Besides, both of *GSA* and *BA* have approximately acquired the same results.

Conclusion

In this study, we have proposed *IWO* algorithm which is one of metaheuristic optimization algorithms to solve the *CD* problem. In order to analyze comparatively, *BA* and *GSA* have also been used in the experimental studies. For experimental studies, four different social networks have been utilized and these networks have been run under the same conditions. When the results obtained by experiments have been examined, it is cleary observed that *IWO* has been much more successful than the other methods in terms of both of solution quality and execution time. Additionally, *IWO* has achieved far superior success for the network called *Dutch College* with the largest number of edges. Therefore, it can be concluded that this algorithm can lead to much better results than the other algorithms in large networks.

References

Attea, B. A., Hariz, W. A., & Abdulhalim, M. F. (2016). Improving the performance of evolutionary multi-objective co-clustering models for community detection in complex social networks. *Swarm and Evolutionary Computation*, *26*, 137-156. doi:10.1016/j.swevo.2015.09.003

Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E, 70*(6). doi:ARTN 06611110.1103/PhysRevE.70.066111

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America, 99*(12), 7821-7826. doi:10.1073/pnas.122653799

Leskovec, J., Lang, K. J., & Mahoney, M. (2010). *Empirical comparison of algorithms for network community detection.* Paper presented at the Proceedings of the 19th international conference on World wide web.

Li, J. W., & Song, Y. L. (2013). Community detection in complex networks using extended compact genetic algorithm. *Soft Computing, 17*(6), 925-937. doi:10.1007/s00500-012-0942-1

Makagon, M. M., McCowan, B., & Mench, J. A. (2012). How can social network analysis contribute to social behavior research in applied ethology? *Applied animal behaviour science, 138*(3), 152-161.

Mehrabian, A. R., & Lucas, C. (2006). A novel numerical optimization algorithm inspired from weed colonization. *Ecological informatics, 1*(4), 355-366.

Newman, M. E. J. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E, 69*(6). doi:ARTN 066133
10.1103/PhysRevE.69.066133

Papadopoulos, S., Kompatsiaris, Y., Vakali, A., & Spyridonos, P. (2012). Community detection in Social Media Performance and application considerations. *Data Mining and Knowledge Discovery, 24*(3), 515-554. doi:10.1007/s10618-011-0224-z

Pons, P., & Latapy, M. (2005). Computing communities in large networks using random walks. *Computer and Information Sciences - Iscis 2005, Proceedings, 3733*, 284-293.

Shi, Z. W., Liu, Y., & Liang, J. J. (2009). PSO-based Community Detection in Complex Networks. *2009 Second International Symposium on Knowledge Acquisition and Modeling: Kam 2009, Vol 3*, 114-+. doi:10.1109/Kam.2009.195

Steinhaeuser, K., & Chawla, N. V. (2008). Community detection in a large real-world social network. *Social Computing, Behavioral Modeling and Prediction*, 168-175. doi:Doi 10.1007/978-0-387-77672-9_19

Zhou, X., Liu, Y. H., Zhang, J. D., Liu, T. M., & Zhang, D. (2015). An ant colony based algorithm for overlapping community detection in complex networks. *Physica a-Statistical Mechanics and Its Applications, 427*, 289-301. doi:10.1016/j.physa.2015.02.020

Zhou, Y., Chen, H., & Zhou, G. (2014). Invasive weed optimization algorithm for optimization no-idle flow shop scheduling problem. *Neurocomputing, 137*, 285-292.