Community Discovery Algorithm Based on Clustering and Genetic Optimization

Befikadu Birtukan Sieyum

[1]Tianjin University of Technology and Education, School of Information Technology and Education, 1310 Da gu nan Road, He xi District, Tianjin, P.R. China

Abstract: ***Community discovery algorithm is recently active area of scientific research and a study of in real world networks such as, computer network, social networks. Social network is a complex network of includes community groups, that have relationship between people in common identity, location, interests, occupations etc. It is used to have better standard community structure in complex network. This study proposed that a combination of clustering which is specified in k-means algorithm and genetic algorithm. In community discovery research area, there are many methods to solve a problem, because of this article depends on overlapping community study used the clique percolation method (CPM) to add in both algorithm that gives a better result in previous works. The study improves to have well structure community; quality of the relationship between two nodes satisfied and accurate relationship between each network in community.***

Keywords: Community discovery, Clustering algorithm, Genetic optimization, k means clustering

1. Introduction

Considering community discovery algorithm is a study of human life and their activities communication. It's important studying communication between living things, because within communication happen relationship creates a network. Effective usage of those relationship in network plays important role for every one growth and development. The community discovering roughly divided in to two kind, overlapping community and non-overlapping community, it's named also the disjoint community.

Overlapping community called also nodes with many communities, the condition of nodes in the network to be a part of another community which means belongs to both the association and others. In many social and information networks, vertices that has more connections inside the set and outside the set these communities naturally overlap. The overlapping node is the vertex that belongs to more than one community. It's NP hard complexity problem and the graph also undirected and also un weighted.

There is some algorithm research based on overlapping community discovery on undirected networks. This article analyses overlapped community have a better community structure, gives to provide an effective services and better community structures.

The proposed algorithm based on clustering and genetic optimization, on previous work research results inspired by definition of poor connection in complex network to solve between two individual's relationship problem and increase the population diversity. In complex network community discovery algorithm methods based on genetic algorithm with clustering combination have very effective, helpful to get better result [1]. The problem of community discovery in complex network is more related to in graph theory. Because the representation of network graph is complicated not easy to conduct numerical calculations in mathematics. To describe the data model, have transferred in matrix form. For this kind of problem, the most common used matrix is adjacency matrix model but, the problem of community is completely not applicable in this matrix; so that the experts discovered Laplacian matrixes [2].

Graven and Newman (GN) is one of the community discovery algorithms that analyze the edges betweenness and finding community structure, detect the communities by progressively removes edges from the original network [3] it can discover cluster overlap communities. In the field of overlapping community discovery studies overlapping agency in 2005 [4]. The Label propagation algorithm (LPA) is assigns the labels from previously a labeled data points at the start of the algorithm. It helps to have changes unique label for each node and densely connected group reach a common label quickly [6].

Clique based method is every node in connected to every other node, it can analysis of how the overlap network gives over lapping community. Clique percolation method (CPM) mostly use in overlapping community discovery researches, examine the network finding all cliques in NP problem, its raised too much selection of the parameter clique in the past research [7]. Researchers [9], [10] conduct to mix k-means together with genetic algorithm program to reach obtaining optimum resolution and to return out of the native optimum. in addition, a settled selection regarding clusters or low-level formatting over population in imitation of minimizing the run time over performance could also be viewed among that paper[11]. In many combinations of algorithm work mostly the results are better than previous works.

This paper, the methodology of community discovery with genetic optimization present in section 2. In section 3 discus about how to implement the proposed algorithms, and how to perform genetic operators. Section 4 shows the result analysis, and conclusion and future work include in section 5.

2. Methodology of Community Discovery

In this paper the genetic algorithm is the main structure algorithm of proposed algorithm, so this section focuses on how genetics can improve the k means algorithm and the specific implementation process. Reading the data as a community detecting used for undirected and unweighted graph initialization CPM algorithm is a method that uses for overlapping community detection by assign a number of k cliques. For using this algorithm in initialization, helpsto select easily, and to detect which area have densely or overlapped community.

In the genetic algorithm the chromosome a population consisting of some solutions where the population size the number of solutions, each solution is called individual. Each individual solution has a chromosome, it represented a set of parameters defines individual. Each chromosome has genes with two representation those are one is genotype is the set of genes representing the chromosome; the other is phenotype, it's the actual physical representation of the chromosome. A collection of many solution chromosome is called population.

After chromosome there is the evaluation of fitness function calculation, the GA has its own unique advantages and it only needs to be evaluated according to the fitness function for optimum results. Candidate solutions, without relying on too much other information, and only determine the subsequent legacy based on the fitness function value Pass the operation. The fitness function should be satisfied the requirement being clearly defined that implemented efficiently, generate intuitive results and it should also quantitatively measure how fit a given solution is in solving the problem. on the study of community network, the fitness function calculated using modularity is a good one. When the overall nodes pare sum, it gives the modularity (Q) equation:

$$Q = \frac{1}{2m}\sum_{n_1,n_2}\left[A\eta_1 n_2 \frac{d_{n_1}\, d_{n2}}{2m}\right]\frac{C_{n_1}c_{n2}+1}{2} \qquad (1)$$

It's good for partitioning two communities only in addition it can generalized for partitioning a community in to different group communities to make possible approach for identifying multiple communities in a network.

$$Q = \sum_{l=1}^{c}\left(e_{ij} - a_i^2\right) \qquad (2)$$

where e_{ij} is the fraction of edges with one end vertices in community i and the other in community j:

$e_{ij} = \sum_{n1n2} \frac{G_{n,n2}}{k_n m} 1_{n1} \in c_i\ 1_{n2} \in c_j$ and a_i is the fraction of ends of edges that are attached to vertices in community i:

$$a_i = \frac{d_i}{k_n m}.$$

Process the genetic operator (selection, crossover and mutation) are used to get the optimum results. The selection operator use fitness function select the parent chromosome and prepare for cross over operation. Crossover is made in one-point crossover. The mutation will apply on randomly flipped the offspring value to finalize the new child.

3. Design and Implementation

This paper proposes the community clustering k means genetic algorithm (CCK-GA). It is a combination of genetic algorithms and clustering algorithms with over lapping community algorithm in CPM.

3.1 Initialization

Population initialization: proposed algorithm when on discussing the overlapping community the method of CPM initialization. The encoding method currently applied to genetic algorithm have real-value based encoding genetic algorithm which means it represents a gene in terms of value or symbol or string. For example, let's say we have the network shows bellow in figure1(a). There is undirected a complete graph and, in this graph, let use if the cliques take k = 3 because of my k value chooses the graph that has three nodes connected in group, by using a graph network. It can form seven sub groups, those are: 2, 3, 3; 2, 3, 4; 2, 4, 5; 3, 4, 5; 4, 5, 7; 7, 8, 10 and 8, 9, 10 nodes identify also with index value.

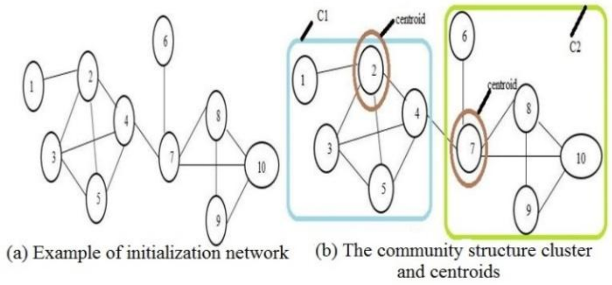
Generate chromosome: In the first step of generation and representation is the selection of k clique group value select two k cliques group result. in step two it will form 5 number of index chromosomes counting start from zero. After step two it can make the reading which nodes are formed and if there are duplicate values of gene in node representation can remove randomly one of it and replace by zero if not it will continue for the next step.

3.2 Fitness evaluation

k means algorithm: in this step use adjacency matrix for calculating their degrees and identifies k number of centroids with in chromosome length 6 cluster to have a partition and structure, then allocates every data point to the nearest cluster, while keeping the centroids length small as possible. Using chromosome result to cluster each similar community in one group uses count the node number that have in chromosome, analyses the distance has highest number of degrees. After this process that identifies the community structure cluster and centroids, as shown in figure 1(b).

Modularity in fitness function: it's necessary to analyze quality of community generation in each chromosome nodes, by using chromosome information it can calculate modularity to have best population based on fitness function in genetic algorithm by using the equations that mentions in section 2.

Selection Operator: it offers preference to higher people, permitting to die their genes to the subsequent generation. The goodness of every individual depends on its fitness. Fitness is also determined associate objective perform or by a subjective judgement. Parent (a) is 2, 0, 5, 4, 0, 7 and parent (b) is 3, 4, 5, 7, 8, 10.



(a) Example of initialization network (b) The community structure cluster and centroids

Figure 1: Example of community discovery network

Crossover Operator: The new off springs created pairing are place into the subsequent generation of the population. By recombining parts of excellent people, this method is probably going to make even higher people. The process of crossover is one-point crossover by using randomly method, to have the crossover point randomly for each pair of parents to be mated, a crossover point is chosen at random from within the genes. Use the probability point and select one crossover points have and exchange from parent (a) with parent (b). After processing the crossover, the new offspring may be having duplication node n this situation change the duplicated node randomly replace one node with zero, otherwise gust use for next step mutation process.

Mutation Operator: With some low chance, some of the new people can have a number of their bits flipped. Its purpose is to keep up diversity at intervals the population and inhibit premature convergence. only using randomly method change some characteristics of new children chromosome by flipping method. The algorithm terminates if the population converged fitness function result, does not produce different offspring from the previous generation.

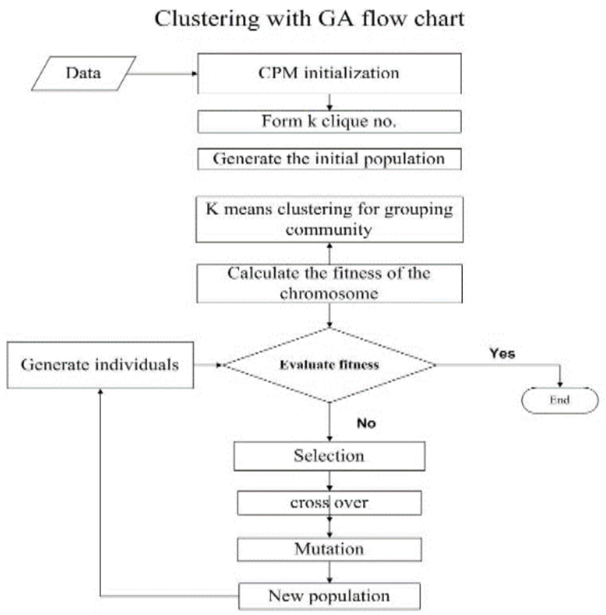

Clustering with GA flow chart
Data
CPM initialization
Form k clique no.
Generate the initial population
K means clustering for grouping community
Calculate the fitness of the chromosome
Generate individuals
Evaluate fitness
Yes
End
No
Selection
cross over
Mutation
New population

Figure 2: Flow chart of combining clustering with GA

The proposed combination algorithm flow chart will show on figure 2, it leads that the increment of service providers in network community and reduce running time to get a good service; and also, to know the algorithm of best than previous works improved genetic in clustering analysis different techniques.

4. Result analysis

4.1 Datasets

To evaluate the performance of the planned algorithmic rule with totally different universe networks 3 categorical datasets from network UCI Machine Learning Repository. Those datasets are social networks and has been widely used in community discovery researchers there are Football, Dolphin and Pol-books;

Dolphins: The bottle nose dolphin network was rumored in [12]. It represents a network of dolphins living in uncertain Sound. The nodes of the network represent dolphins, whereas the perimeters represent the connection between 2 dolphins.

Football: The yanked faculty soccer Network was derived by Newman [13] in 2004. The nodes of this network represent soccer groups, whereas the perimeters represent the matches between the groups. The groups were classified into 12 totally different teams, but 8 freelance groups.

Pol-books: this is often the social network folk's politics books. The nodes of network represent books concerning us politics, whereas a position between 2 books represents that the books were purchased along by a peremptory. Books are categorized in 3 categories: liberal; neutral; and conservative.

After performing experiments on various networks with different parameter settings, initialized the final parameters with the following values: population size = 200; crossover rate (Pc= 0.4); mutation rate (Pm = 0.01); maximum generation (maxgen= 25).

4.2 Results

On the applying of the CCK-GA algorithm based on those data have analyzed in python language with the criteria mentioned above. In order to run the experimental algorithm on the mining of complex network communities, the result running after 10 times apply and clusters into different groups according to centroid is shown in figure 3.

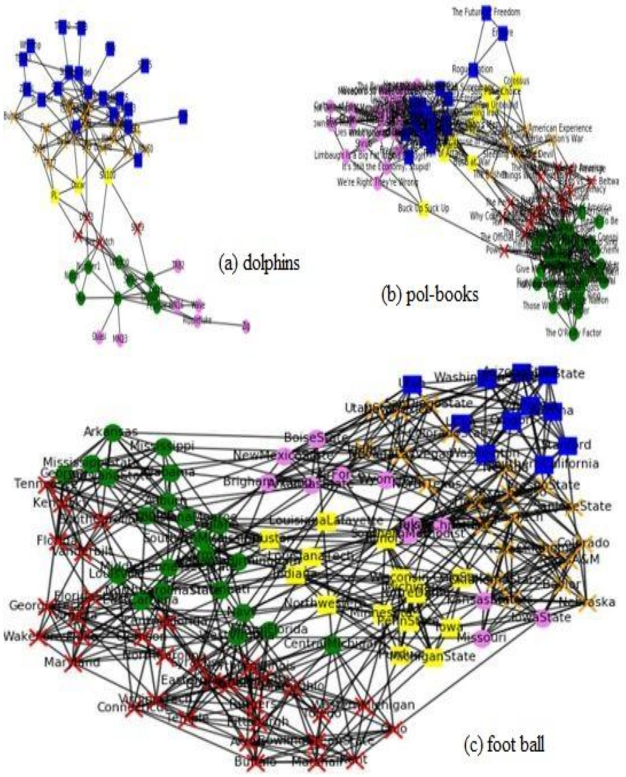


Figure 3: The graph of networks by using CCK-GA

This figure shows the main division communities' structure by using a partition of 6 centroids based on chromosome length.

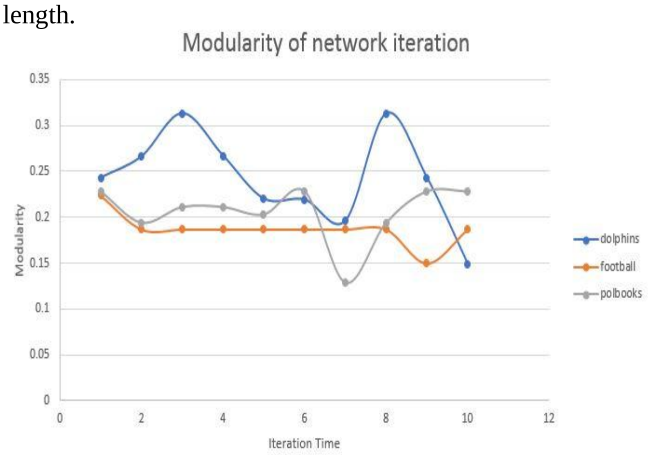


Figure 4: The Modularity of network with iteration time

In figure 4 have the result of the best population in Dolphins seen on the generation of 1st with 10th iteration time, Football seen on the generation of 2nd with 9th iteration time, Pol-book seen on the generation of 13th with 7th iteration time, and the k means get good structure for each data.

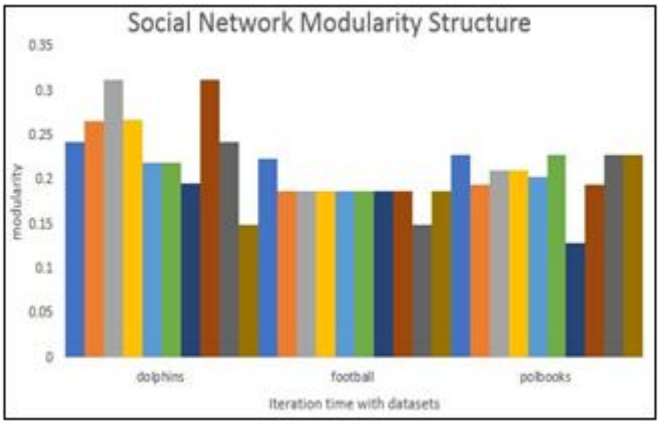


Figure 5: The Social network modularity structure

In figure 5 that shows the structure of for all network dataset with the iteration time that runs in 10 times are shown with different colours describe that each iteration time from 1to 10 and the generation modularity structure included.

To compare the proposed algorithm modularity takes the previous work and from information map. The optimum result shows based on genetic optimization the modularity of all population maximization is shown in table1.

Table 1: The optimal result proposed algorithm of Q

Networks	*Info map*		*CCK-GA*	
	Q_{max}	Q_{avg}	Q_{max}	Q_{avg}
Dolphins	0.528	0.524	0.313	0.267
Football	0.601	0.601	0.224	0.195
Pol-books	0.523	0.523	0.228	0.203

Table 1 shows combining k means clustering and genetic optimization with the four data network community cluster and maximum modularity results for each network and have improved the genetic algorithm clustering community thought in the paper is obtained in function optimization. Good results, but more needs detailed research on complex network issues, improve algorithm CCK-GA Optimized efficiency in mining complex network communities specially to improve in the small network nodes result.

5. Conclusion and Future work

Studying of community discovery within the past few years, it helps to own a far better structure and effective time and area quality results. This helps to facilitate real human life within the easiest method furthermore studies to get a far better solution for each community networks.

Discovering a good community structure; The proof of the experimental analysis shows that gives solutions of higher quality with equivalent procedure resources. The results that statistically full and the parts represent with k means based on network degree and genetic improvement contribute to the determined performance. To apply different comparation in community algorithm methods and proposed algorithm it needs more implementation will include as a future work.

References

[1] He Dong-xiao, Zhou Xu, Wang Zuo, Zhou Chun-Guang, Wang zhe, Jin Di. communitymining in complex networks-clustering combination based genetic algorithm ActaAutomaticaSinica, 2010,36(8):1160-1170

[2] Chaiken, S. and Kleitman, D. "Matrix Tree Theorems" [J], jornal of combinational theory, series A24 (3): (1978)377-381.

[3] M.E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. Phy. Rev. E69, pp. 026113, (2004)

[4] Steve Gregory, an algorithm to Find Overlapping Community Structure in Networks [J], Knowledge Discovery in Databases: PKDD 2007: 91.

[5] A. Clauset, M. E. J. Newman, C. Moore. Finding community Structure in very large networks [J]. Physical Review E, 2004, 70(6):066111.

[6] Blondel V.D., Guillaume J.L.,Lambiotte R, et al, Fast unfolding of commun in large networks[J] Journal of Statistical Mechanics: Theory and Experiment, 2008, 2008(10) P10008

[7] Palla G., Derenvi I., Farkas I. et al. Uncovering the overlapping community structure of complex networks in nature and society [J]. Nature, 2005, 435(7043): 814-818.

[8] Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D. Defining and identifying communities in networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(9): 2658-2663.

[9] Kernighan, B. W.; Lin, Shen (1970). "An efficient heuristic procedure for partitioning graphs Bell SystemTechnical Journal, **49**: 291–307, doi:10.1002/j.1538-7305. 1970. tb01770.x

[10] Palla, Gergely (2005), Uncovering the overlapping community structure of complex networks in nature and society, 435(7043):814-818. arXiv: physics/0506133, bib code: 2005 natur, 435. 814p. dio: 10.1038/nature03607. PMID 15944704

[11] Stanley Wasserman, Katherine Faust, 1994. Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press.

[12] D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten, S.M. Dawson, the bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations, Behav. Ecol. Sociobiol. 54 (4) (2003) 396–405.

[13] Mc Roberts, O.M. (2001, January/February). Black Churches, community and development. Shelter force Online. Washington, DC: Author. at nhi.org

Author Profile

Befikadu Birtukan Sieyum received the Bachelor Degree, Hard ware and Network Engineering Service Management from Entoto Polytechnique college, Addis Ababa, Ethiopia in 2014. Currently studding M.E on Applied Computer Technology in Tianjin University of Technology and Education.