Network Community Structure Clustering Algorithm Based on the Genetic Theory

Nan Lu, Yuanyuan Jin, and Lei Qin

Abstract—This paper proposes the idea of applying a clustering ensemble based genetic algorithm in the area of complex social network mining. The algorithm introduces clustering ensemble into the crossover operator and employs the clustering information of the parents to generate new individuals, which avoids the problems that caused by simply exchanging string between crossover operators without consider the contents. In population generation, Markov random walk strategy is employed to maintain the diversity of the individuals as well as the clustering accuracy. The algorithm also uses a local searching mechanism in crossover operators to reduce the searching space and improve the speed of convergence. Comparing with existing mining algorithms in social network, the proposed algorithm is more effective proved by experiments in both simulation and real world social networks.

Index Terms—Community structure, complex network, genetic algorithm, clustering ensemble, data mining.

I. INTRODUCTION

Complex networks are associated network which representing real world situation, each of which consists of node (entity)-sets and edge (associations between entities)-set. The metabolism network and gene regulation network in biology; Internet, WWW, e-mail system in application; Interpersonal network in social science make good examples of the complex networks. The study of complex network theory attracts researchers from different areas such as: computer, physics, math, biology, social and complexity science and therefore becomes a hot spot in the interdisciplinary researches[1]-[3].The study of various types of networks (such as biological networks, science and technology network, exhibits basic statistical characteristics of small world [2], and scale-free [3].

Community structure is an important characteristic of complex network, Community is a group of nodes similar with each other but different from other nodes [4], [5]. Nodes in the same community are intensively connected, while the connections between nodes within different communities are sparse [6], [7]. Similar to clustering in data mining, community structure mining is employed to find the important clustering characteristic of the community. It is also meaningful in analysis of the structure and function of the complex network topology as well as behavior prediction. Community structure mining has been widely applied in terrorist recognition, organization structure management, metabolic path prediction, protein interaction network analysis, Web community mining, search engine, etc.

Currently, the researches of complex network community mining are mainly divided into two types: one is community modularity function based hierarchical clustering method, which is aiming at clustering community structure in networks. Examples of these types are: Girvan Newman (GN) algorithm based on the decomposition [4], Label Propagation (LP) [8], Propinquity Dynamics (DP) [9], and Similarity Dynamics and so on. The other one is by transforming the community mining problem into optimization problem; such as: Kernighan Lin (KL) algorithm using heuristic and optimize strategy. It finds the maximum objective function value by utilizing gain function to divided network into communities. However, the clustering of predefined optimization objective function is considered to be a NPC problem[10], and therefore can not be solved by exhaustively search. Genetic algorithm (GA) has been proved to be a very effective method to solve NP problems with guaranteed quality solution with a reduced time complexity. But most existing GA community mining algorithms [11], [12] are still weak in optimization power because of the slow convergence rate. The main problem lies in the direct application of traditional cross—operators in the complex network community mining context, which will make these methods to avoid or provide less effective crossover operators.

This paper presents a Clustering Combination on Complex Networks (CCCN) algorithm based on genetic theory to mining Complex network community structure. The algorithm propose the idea of clustering by cross-operators in genetic evolution process, which avoid the problem caused by traditional crossover operator without considering the clustering content and simply exchanging characters block. By this method, the desirable characteristics from the previous generation will be effectively inherited by the new individuals of next generation, which will amplify the global search ability of the crossover operators. The contributions of the CCN algorithm are summarized in the following aspects: 1) proposes initial population generating algorithm based on Markov Random Walk, which will produce diversified initial individuals with a certain clustering accuracy. 2) the initial population generated by the algorithm help in building up an optimal initial solution in a certain extend. The initial population also collaborates with the crossed operators to improve the optimal searching ability and speed up the search process. 3) Compared with the traditional
random mutation operation, the algorithm uses local search on mutation nodes and most of their neighbor nodes in the community to complete mutation operation. It can effectively reduce the search space, search efforts and speeding up the convergence of genetic algorithm. The experimental results show that the CCCN algorithm's effectiveness.

II. RELATED PROBLEM DESCRIPTION

A. MCSCN (Mining Community Structure in Complex Networks)

Complex network can be modeled as a graph \( G = (V, E) \), where \( V \) is node set of the network and \( E \) is the set of edges. A community of a network is a set of nodes having the characteristic of condense edge linkages within the community, and loose linkage between communities. The community mining problem is to identify community structures from various complex networks.

B. CM (Community Modularity)

Community modularity is a evaluation function introduced by Newman \( ^{[4]} \) to measure the quality of community structure identification, also known as \( Q \) function. Assume that a complex network is divided into \( k \) communities, a \( k \times k \) symmetric matrix \( E = (e_{ij}) \) is defined, where \( e_{ij} \) is the number of edges that connecting two different nodes from the \( i \)-th community and \( j \)-th community divided by the total number of edges. Then modularity \( (Q) \) expressed as:

\[
Q = \sum_{ij} (e_{ij} - a_i^2) = Tre - \| e^2 \| \tag{1}
\]

In formula 1, \( Tre = \sum_i e_{ii} \) is the sum of each element on the diagonal of matrix \( E \), it expresses the proportion of the number of edges from each node in a certain community to the total number of edges. \( a_i = \sum_j e_{ij} \) is the sum of each element of each line, it represents the proportion of the edge number of the \( i \)-th community nodes to the total number of edges. The weaker the community structure, the smaller the \( Q \) value will be. Generally speaking, if the \( Q \) value is more than 0.3, the network is considered to be a significant community structure. Although \( Q \) function has the resolution limit, most current studies consent that it is the standard of the quality of network clustering results. The proposed algorithm CCCN also uses \( Q \) function as a objective function.

C. NCM (Node Classification Mistake Rate)

To illustrate the accuracy of clustering, node classification mistake rate is employed. NCM corrects those nodes that have obvious clustering mistake. If the mistake nodes exceed a certain threshold, the results will converge to a local extreme value, which is incorrect. Consider the concept of community, the possibility of the node with its neighbor node to be in one community should be maximized. Through the analysis of each node and its neighbor nodes, a category \( ID \) value is used to improve the accuracy of the classification results. We define \( NCM (i) \) to represent node \( i \) current classification mistake rate:

\[
NCM (i) = \frac{\sum_{j \neq i} f(i, j)}{\text{deg}(i)} \tag{2}
\]

Among the formula above,

\[
f(i, j) = \begin{cases} 1 & \text{classID}(i) \neq \text{classID}(j) \\ 0 & \text{classID}(i) = \text{classID}(j) \end{cases} \tag{3}
\]

where \( \text{deg}(i) \) is the degree of node \( i \), \( E \) is the set of edge in the network, \( \text{classID}(i) \) for node \( i \) category. If the \( NCM (i) \) is greater than a threshold, the classification of the node is considered to be incorrect.

III. GENETIC ALGORITHM

In genetic algorithm (GA), the optimal solution is achieved by genetic operation on chromosome individuals which with the bit string structure. Each chromosome individual is a candidate solution. Individuals are produced through crossover and mutation (renewing and restructuring) operations to improve the overall quality of the individual evaluate function. The basic process of genetic algorithm includes: initialization, individual evaluation, population evolution (selection, crossover and mutation) and termination inspection.

A. Construction Complex Network

First of all, the similarity between each pair of nodes is calculated from the original data to construct a similarity matrix. The elements in similarity matrix is greater than the threshold value will be labeled as 1, others are labeled with 0 (that is, removing weak connected edge). This generated 0, 1 matrix is similar to the adjacency matrix of the network and will reduce the calculation complexity.

B. Genetic Coding

Community structure mining based on GA in its search process need to maintain a population generated by the classification result from different network society. In [13], the gene bit character adjacent string coding LAR (Locus -based Adjacency Representation) was applied to the community mining algorithm based on GA. The operation of Character string coding mode is simple, but it’s difficult to cross, hard to decoding, not efficient and doesn't fit in to the fitness function evaluation. We adopt a character string coding form to effectively solve the problem in genetic crossover and decoding. If a network contains seven nodes \( \{v1, v2, v3, v4, v5, v6, v7\} \), if the character string coding chromosome is \( \{1, 2, 1, 2, 1, 2, 2\} \), then it predict that the network node \( \{v1, v3, v6\} \) belong to a network community, node \( \{v2, v4, v5, v7\} \) belong to another network community.

C. Population Initialization

To adapt to the complex network clustering ensemble requirements and taking all advantages of the global search ability of multiple individuals’ crossover operator. Only when the individual has a certain accuracy and strong diversity, can make full use of network clustering ensemble effect. This paper adopts a random walk method based on MARKOV chain to initialize population.
If the individual conducts a random walk along the edge of a network (non-weight, undirected network), the random walk should follow the transition probability to choose the next arrived node. If the individual current position node is \( i \), the next step to neighbor node \( j \) probability \( P(i, j) \) can be expressed as \( P(i, j) = 1/d_i \), including \( d_i \) is degree of node \( i \). Obviously, the sum of transition probability of the nodes in the network to node \( i \) is 1, namely \( \sum P(i, j) = 1 \).

Let the \( P_t^i(i) \) be the probability of a individual starting from any node \( i \) and get to destination node \( t \) after \( al \) step random walk, it can be expressed as:

\[
P_t^i(i) = \sum_{<i,j>} P(i,j) \times P_t^i(j), i \neq t (1 \leq i, t \leq n)
\]  

(4)

The \( <i,j> \) expresses edge connecting node \( i \) and node \( j \). The vector component of probability value every node in the network random walk \( l \) step arrive at the destination node \( t \) called \( r' \) s \( l \) step transition probability distribution, express as:

\[
P_t^l(i) = \sum_{<i,j>} P(i,j) \times P_t^l(j), i \neq t (1 \leq i, t \leq n)
\]  

(5)

where \( n \) is the number of node in the network. Random walk can be regarded as the stochastic process based on MARKOV connection attributes. Due to the density of the connection in the community is obviously higher than the connections between community and the initial node of individuals in the network and the selected destination node is located in the same community, there should be more path choice which experience \( l \) step migration finally arrive at the destination node \( l \) value can’t be too big. On the contrary, the individual finally arrive at the destination node probability will be smaller. \( P_t^l(i) \) should meet:

\[
\big\langle \forall ic_t \forall j \in C_i \big\{P_t^l(i) > P_t^l(j)\big\}, 1 \leq i, j, t \leq n
\]  

(6)

where \( C_i \) is that the community destination node \( t \) located.

Thus it can be seen in the premise of guarantee the accuracy the MARKOV random walk method generated initial population, with strong diversity, very suitable for clustering ensemble. And its generated initial population at a certain extent makes the initial solution space close to the optimal solution space, which can accelerate the convergence process of mining community structure.

Initialization process used an integer array \( a[n] \) representing the solution space \( n \) is node number). Array for each data type \( (classID) \). Both \( a_i \) represents \( i \)-th node category number. Array \( a \) is regarded as a population of genetic algorithm (that is, a solution). \( N \) is for population length. Each population represents different classification information.

D. Clustering Crossover Operator

Crossover operation is the most important genetic operation of genetic algorithm, in order to get better clustering effect, in view of the clustering algorithm in the design of the data mining is the strategy used to solve which balance multiple clustering results. We will with the genetic algorithm crossover operator through mixing different population information to produce new population, and the genetic algorithm for the two traditional population crossover operations is expanded to multiple population crossover operation, and then get multi-population crossover operator based on clustering ensemble.

Definition 1. (edge association) Edge association is use \( N \) community classification clustering information, through clustering ensemble method, presents a new metrics judge whether a edge in community internal.

Known have \( N \) network community division, take either edge in the community \( <u, v> \), the relational \( C_{uv} \) defined as:

\[
c_{uv} = t_{uv} / N
\]  

(8)

which \( T_{uv} \) express the number of edge \( <u,v> \) in \( N \) community division located in the community internal. \( C_{uv} \) can also be interpreted as probability of edge \( <u,v> \) located in community internal. Obviously, the greater the \( C_{uv} \) edge \( <u,v> \) is tending to be in a community internal. On the contrary, edge \( <u,v> \) is more likely exist as a associate between community. So can think, if network community structure is reasonable, the correlation degree of edge in the community internal should be greater than correlation degree of edge between communities.

Definition 2. (edge topological similarity) uses network topology to calculate the edge topological similarity. A network \( G=(V,E) \), \( <u,v> \in E \), define edge topological similarity \( Sim(<u,v>) \) as:

\[
Sim(<u,v>) = \frac{|S(u) \cap S(v)|}{\sqrt{|S(u)| \times |S(v)|}}
\]  

(7)

The \( S(u) = \{v \in V | <u,v> \in E\} \cup \{u\} \) express neighbor set of node u. The practical significance of the definition is expressed as: the bigger the circle of friends two people sharing in social network, and the familiar the two men also may. So if the bigger the structural similarity of a edge suggests that its two vertices more similar, that is to say the edge more easier located in network community internal.

From that, 1) based on the clustering ensemble crossover operator can effectively avoid the problem brought by the traditional crossover operator does not consider clustering content, simple exchange character block. 2) make reasonable use of effective clustering information provided by father population, produce good new population, thus to strengthen the global search ability of crossover operator. 3) initial population generating algorithm based on the random walk provides the initial population suitable for clustering ensemble, make clustering ensemble effect can be brought into full play, so as to further enhance the global optimization ability of the crossover operator.

E. Variation and Selection Operator

As a local search operator of genetic algorithm, and the function of mutation operator is on the premise of given the mutation rate, implementation variation for each gene of chromosomes. Using local search mechanism, forced variation node and its most neighbor nodes take the
variation operation in the same community. If community contain variation neighbor number more than one, from these random choice one as the new community the node belong to. Compared with the traditional variation method, 1) the targeted to reduce the search space problem, reduce the insignificance search, accelerating genetic algorithm convergence. 2) effective solving other genetic operator produce individual node misplacement problem.

The selection operator is the global search operator of genetic algorithm. To keep the optimal individual and speed up the algorithm convergence speed of every generation, we adopt the μ + λ selection strategy of combination optimization. μ + λ strategy is from the father population and new population produced through the crossover and mutation chooses the population with the biggest fitness scale as progeny population.

**F. Clustering Correction**

In order to increase the accuracy of clustering, must correct those nodes which may be obvious division error. If the error nodes is too much, will lead to the results converge to a local extreme value, so as to get a wrong division. We adopt the step through the node classification error rate NCM for error correction.

Firstly, according to the structure sequence of the node number analysis all nodes NCM (i) value, if the NCM (i) bigger than a given threshold value, then the division of the node present obvious error condition.(2) find the node’s neighbor nodes classID, appear the most times classID replace the original classID. Through clustering error correction step, greatly solve the community structure clustering error.

**IV. CCN ALGORITHMS**

**A. Parameters Set**

CCCN the parameters and value as is shown in table 1. Among them, the parameters Pop_scale, Max_gen_times, non_improve_maxgen, alpha, Pmutation_rate, tournament_scale_size is the basic parameters of the genetic algorithm, their value are chosen according to the experience. Random walk steps l is taken from [14], individual number M used for cross operation is the parameters of many individual crossover operator based on the clustering ensemble proposed in this paper.

<table>
<thead>
<tr>
<th>Parameter name</th>
<th>Parameter meaning</th>
<th>Parameter value</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>new individual accounts for the proportion of the population scale</td>
<td>0.15</td>
</tr>
<tr>
<td>Pop_scale</td>
<td>Population scale</td>
<td>10</td>
</tr>
<tr>
<td>Max_gen_times</td>
<td>The limited of iteration numberl</td>
<td>500</td>
</tr>
<tr>
<td>Non_improve_maxgen</td>
<td>The optimal individual unchanged maximum iterative number limit</td>
<td>100</td>
</tr>
<tr>
<td>n</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pmutation_rate</td>
<td>Mutation rate</td>
<td>0.1</td>
</tr>
<tr>
<td>Cross_num</td>
<td>The individual used for cross operate</td>
<td>20</td>
</tr>
<tr>
<td>tournament_scale</td>
<td>The scale number in mutation and choice method</td>
<td>20</td>
</tr>
<tr>
<td>Walk_steps</td>
<td>The step of random walk</td>
<td>10</td>
</tr>
</tbody>
</table>

**B. Algorithm Description**

1) initialize population scale Pop_scale
2) according to the Pop_scale population scale, use the markov random walk method which describe in 3.3 to generate initial population P = {P1, P2,..., Pn}.
3) initialize each new population Pnew = 0, new = 1, 2, ..., n.
4) for i = 1, to α * Pop_scale (α is new individual accounts for the proportion of the population scale (0 < α < 1)).
5) use the description in 3.4 which through multi-individual crossover operator based on clustering of many producing a new individual I cross.
6) use the method in 3.5 to apply mutation operator based on the local search on the I cross to produce a new individual i
7) Pnew←Pnew ∪ Imut, create new population.
8) return to (4) and use the cross operator to generate new individual
9) P←P ∪ Pu, form a new population number
10) from P choose a Pop_scale individual with the largest fitness value as the next generation population.
11) until meet the conditions, the algorithm termination.

The highest time complexity in CCCN algorithm is: step 1 (generating initial population) generating initial population time complex is O (n2) and step 5 (many individual crossover operations based on clustering ensemble)the time complexity of many individual crossover operations based on clustering ensemble is O (n2). The algorithms at most perform the cross operation Max_gen_times×α×Pop_scale times, because generally think of maximum iterative times and population scale of genetic algorithm are constant, operation time of other steps is less than O (n2), so the time complexity of the whole algorithm is O (n2).

**V. THE EXPERIMENT AND RESULT ANALYSIS**

**A. Simulation Network Experiment**

By using the computer generated simulation network to test mining community structure algorithm performance usually become a referenced method. Generated simulation network has the following features: each network contains 288 nodes, six communities, and each community contains 48 nodes. Each node and its community internal node connection number is zin, and other community node connection number is zout, with zin + zout = 16. By increasing zout, the network community structure becomes more and more fuzzy and becomes more challenging for the community mining algorithm. Especially, when zout > 8, the network will be considered as no obvious community structure. We use FVCC (Fraction of vertices classified correctly) [6] as network community mining accuracy measurement method, this method was first used by Newman, et al, has become the most commonly used method for calculation network clustering accuracy. To testing CCCN algorithm performance, we will compare it
with the Girvan Newman (GN) algorithm [4] and Fast Newman (FN) algorithm [15], the two algorithms are classic and with good reference rate. The result of the comparison is shown in Fig. 1 (a).

![Fig. 1(a). CCCN algorithm and the GN, FN algorithm clustering accuracy comparison](image1)

In order to further verify CCCN algorithm validity, the algorithm and the current two classical algorithms: community found extraction algorithm communities and Simulated annealing algorithm SA take comparison experiment, the experimental results are shown in Fig.1 (b).

![Fig. 1(b). CCCN algorithm and the FEC, SA algorithm clustering accuracy comparison](image2)

From Fig. 1(a), the algorithm CCN is better than GN algorithm and FN, and the greater the zout, advantages of the algorithm CCCN more obvious. For example, the algorithm in zout = 7 can give a completely correct community classification results, the algorithm GN can correctly divide the 71.09% node, FN algorithm can correctly divide the 88.28% node. When zout increase to 8, the algorithm will still be able to correctly divide the 99.22% node, the algorithm GN and FN correctly divide the node percentage were 42.19% and 66.41%, which is significantly lower than the CCCN algorithm.

From Fig. 1(b), the CCCN algorithm clustering accuracy is similar to SA algorithm, but it is obviously better than the FEC algorithm, and the greater the zout, the more advantage of the CCCN algorithm is shown. For example, when zout = 7, algorithm FEC and SA clustering accuracy respectively reduce to 78.13% and 84.26%, and the algorithm CCCN can still give a complete right community classification results. When zout = 8, algorithm FEC and SA clustering accuracy reduce to 71.72% and 63.47%, CCCN algorithm clustering accuracy can still maintain at 99.22%.

B. Reality Network Experiment

We through two widely used real network to further verify the effectiveness of the CCCN algorithm. Described in 2.2 Q function as the objective function of CCCN algorithm, so we still choose Q function as the objective function of GN algorithm, FN and SA to comparison experiment.

1) SZUT tennis club network

SZUT is a network built for one university tennis club, the network nodes representing the club members, the edge representative the social communication among the members, which contains 34 nodes and 78 edges. Due to various factors, the club composes of the club management center and club coach center, the two subnets. Based on the network operation algorithm CCCN, Fig. 2 shows division results (node 1 on behalf of the club coach, node 34 on behalf of the club's manager). The right part of the node (diamond nodes and triangle node) represent the members with coach in the same group after segmentation, the left part of the node (circular nodes and square node) represent the members with coach in the same group after segmentation. CCCN algorithm divided the network into 4 parts, respectively shown by four different shape nodes. It can see from this, CCCN algorithm can not only success really dig out the network community structure, but also can further distinguish hidden structure. The Q function value of the algorithm partitioning results is 0.4198, larger than Q function value of the real network division 0.3715. Compared with the other algorithm, the algorithm CCCN get Q value to is obviously superior to the GN algorithm Q 0.4013 and algorithm of FN Q value 0.381, and the result of SA algorithm is the same.

![Fig. 2. CCCN algorithm for Zachary karate club network partitioning results](image3)

2) A dream of red mansions family relation network

A dream of red mansions family relation network is a social network generate by selecting 79 major characters from classic book "a dream of red mansions", with five family (ningguo mansion, RongGuo mansion, wangfu, JiaFu and XueFu) as the basis, connected edge represent the main relationship (here only consider such as parents, brothers and sisters, such as husband and wife relationship). Make α value is 0.75, the population size Pop_scale value from 1 up to 100, the modularity got from the algorithm conccussion with the change of parameters Pop_scale, Pop_scale value equals 60, CCCN algorithm iterative 69 times, can get maximum Q value 0.7402, corresponding to nine community, Shi mansion, wangfu, XueFu was very accurately division out, and the ningguo mansion corresponding to two small community, JiaFu corresponding to three small community, Mrs. Xing and

92
elder brother and his wife corresponding to a single community. To continue algorithm and modularity Q is reduced, when iteration to 72 times, Q value equal to 0.7288, get six community, as shown in Fig.3. At this time, Shi mansion, ningguo mansion, wangfu be accurate division, XueFu only add Mrs Xing and elder brother and his wife the two node, RongGuo mansion corresponding to two communities, and respectively take jia she and JiaZheng as the core. Algorithm to continue iteration to 73 times, Q value reduced to 0.7080, the network was divided into five community, but wangfu and XueFu merge together.

In the same population scale conditions, make α value equal to 0.35, Pop_scale value equal to 30, algorithm running 82 times, achieve the max modularity Q value 0.818. At this time corresponding to 10 community, including history mansion, wangfu, XueFu was a clear division, ningguo mansion corresponding to three community, respectively is jiayan children four generation JiaZhen three generations, YouShi maiden 4 people, including JiaZhen node copied; RongGuo mansion corresponding to three community, respectively is JiaYuan three generations, jia she three generations, JiaZheng three generations, including jia she, JiaZheng nodes be copied. There are several node due to the affinity between family, be divided into error, such as WangXiFeng was born in wangfu, after married within the RongGuo mansion was divided to wangfu, essentially also has a certain rationality.

When Q value reduced to 0.7480, the structure of the division is not clear.

VI. CONCLUSIONS

This paper proposes a genetic algorithm based on clustering ensemble method (CCCN), which can effectively identify community structures within complex network. The algorithm has contributions in following aspects: 1) propose that based on markov random walk thoughts to generate initial population, make the individuals of initial group with certain accuracy and strong diversity, suitable for clustering ensemble. 2) propose multi-individual crossover operator based on clustering ensemble and improve global search ability of the algorithm. 3) local search strategy used in mutation operator to improve the local searching ability of the algorithm.

Further research work includes: 1) algorithm parameters setting, through the more realistic network parameter analysis, more reasonable setting each parameter in the algorithm, make the algorithm can use a shorter time to achieve higher clustering accuracy. 2) improve clustering ensemble method, and to further improve the optimization efficiency of the algorithm.

REFERENCES