Power System Technology
ISSN:1000-3673

Received: 06-06-2024 *Revised: 15-07-2024* *Accepted: 28-08-2024*

Community Detection Using Genetic Algorithm and Improved K-Nearest Neighbor Clustering

Fatemeh Jafari[1], Hamidreza Ghafari[2*]

1,2. Department of Computer Engineering, Ferdows Branch, Islamic Azad University, Ferdows.Iran.

Abstract

Community detection is gathering nodes on graph network's in separate groups that is called community. Detected subgraphs on the network represent the communities in the social network graph. One of the most popular graph clustering algorithms which has become popular due to easy implementation and performance is iterative search algorithm but has problems such as sensitivity to the initial amount and trapping into the local optima trap. At present, k-nearest neighbor clustering is used for clustering data. k-nearest neighbor clustering has sensitivity to neighborhood size k which it's performance severely relies on neighborhood size k. so in this paper first we proposed local mean vector method in k-nearest neighbor clustering for improving and reducing the sensitivity then we use genetic algorithm with improved k-nearest neighbor clustering in order to community detection. Experiments performed on real networks and its subsequent results suggest that the method studied greatly improves the accuracy

Keywords: *k-nearest neighbor clustering, community detection, genetic.*

1. Introduction

The networks are place to exchange a lot of information between peoples. Networks such as Facebook, Twitter and LinkedIn are among the great communication networks. In the simplest form, a social network is a mapping of all relevant edges between the studied vertices. These concepts are often shown in a social network graph in which the nodes are vertices and the lines represent the edges [1].

The mathematical representation tools of social communication in social networks are matrix and graph, the origin of graph theory and its history back to Euler's solution to solve the Seven Bridges of Königsberg problem in 1736 [2]. In a graph, nodes are considered as vertices and the relationship between them represent edges. The community is a subgraph of a graph which number of edges between members in the sub-graph is greater than the number of edges connecting the subgraph to the rest of the graph. The aim of community making is gathering nodes on graph network's in separate groups that is called community [3]. Detected subgraphs on the network represent the communities in the social network graph. The clustering of data sample in existing communities is based on the similarities existing between them. Identifying

Power System Technology
ISSN:1000-3673
Received: 06-06-2024 Revised: 15-07-2024 Accepted: 28-08-2024

and detecting communities in networks plays an important role in a wide range of research fields such as computer science, biology, and sociology. Heretofore, various methods have been developed to detect communities on social networks, which generally include hierarchical methods [4], partition methods[4], edge-based methods, spectral methods[5], modulus maximization-based methods Network[1, 6, 7], probabilistic methods [8] and model-based and optimization-based methods [9, 10] .

In the analytical and hierarchical community detection method, try to find nested clusters recursively in a dense or partitioning state and need to know the similarity matrix or the pattern matrix and method used is based are greedy algorithms and stage optimality [11]. At present, In graph clustering methods, various methods including latent spatial models [5], non-negative matrix factorization [12], block model approximation [13], spectral clustering [14], label back propagation [15] and modular maximization [1, 6, 7] have been used . These models offer different definitions of communities or clustering criteria. According to the application in various conditions, side location models mainly maps network nodes into low-dimensional euclidean space.

The proximity between network connectivity nodes is considered in the new space then nodes are clustered using traditional clustering algorithms such as k-means, k-clique or linkage clustering in a low-dimensional space. Modular maximization models transform the problem of community detection into a modular maximization problem. A modular criterion is usually used as a benchmark for community detection, which by measuring the degree of distribution of nodes measures the degree of community in real networks. These types of algorithms typically use hierarchical clustering techniques to parse the network, which is time-consuming [11].

Like the latent spatial models, non-negative factorization matrix models convert the adjacency matrix of a network to a low-dimensional matrix, then cluster it using k-mean or linkage clustering. Block model approximations consider community detection as a matrix blocking problem which save the index of each node according to their community membership and approximate a given network by a block structure [5].

Modularity maximization models transform the problem of community detection into a Modularity maximization problem. A Modularity criterion is usually used as a measurement for community detection, which measures the strength of a community partition for real networks by taking into account the degree distribution of nodes. These types of algorithms mainly use various hierarchical clustering techniques to partition networks, which is time-consuming [1, 6, 7].

Blondel and et al proposed an idea with application on modularity problems, their algorithm was in heuristic category and based on spreading method and use idea class label propagation models to reduce runtime cost [16]. the heuristic spreading algorithms in compare with other algorithms in area of modularity maximization problems have a good performance in small and

Power System Technology
ISSN:1000-3673
Received: 06-06-2024 Revised: 15-07-2024 Accepted: 28-08-2024

big networks. As an extra, Rosvall and et al proposed a method for community structure which is based on information aware [17]. Their method converts the community detection structure into a information aware coding problem. Furthermore, an information map algorithm of random walks is proposed to solve the optimization problem [18]. There are several studies for research on the performance of existing community detection algorithms. The authors in these papers compared the performance and scalability of these algorithms in real and synthetic networks and the robustness and disadvantage of each algorithm has been evaluated.

In methods based on edges removal, it is assumed that high traffic edges are the bottlenecks of communication between communities, and as a result, the edge with higher path is considered as a cluster edge. By removing these edges from the network, communities are formed in the main graph. One of the common popular algorithms in this area is Greyon and Newman's algorithm [7]. The nodal interfaces have been studied in the past as a centrality criterion and capability of node's influence in network nodes. This criterion describes the effect of node on the information flow between nodes, especially in cases where the information flow on a network is mainly available in the shortest path. In spectral methods, spectral clustering techniques based on the idea of partitioning break down a graph into a subset of vertices which named cut nodes (cutting). Number of produced cuts in these methods is constant and is used to minimize the objective function. The finding optimal cuts problem is formulated as the optimization problem. It has been proved that maximizing the objective function in this problem is a NP-hard problem.

Zhou et al proposed a community detection method based on probability in terms of structure or topics. In the structural method, creates probability of existence of each user in each community using dirichlet parameters and polynominal distribution of users in each community and in the dimensional method, the relationship between communities and topics is recognized, in other words, individuals is clustered based on common topics whereas they may rarely communicate with each other directly [19].

In compression cluster-based methods, a group or cluster from internally view has more compact and denser than other clusters in the network. In the graph theory model, the clustering is named the process of dividing the vertices into groups with a higher density than the edges within the group and a community within a network is considered as a group of dense internal vertices that are connected with each other within the group but are less connected to the external vertices. If a community is more visible, i.e it has denser internal-link among communities and has sparser external-link among communities, so the network clustering of the desired network will be easier and simpler [4, 13].

In model-based methods, if the users gathering in communities is certain, these methods can detect communities such as star and cubic community. With these methods can detect communities in such a way that contain their structural form, for example to create

Power System Technology
ISSN:1000-3673
Received: 06-06-2024 Revised: 15-07-2024 Accepted: 28-08-2024

communities in terms of effective node, the star structure can be used or to have the highest compactness, cubic structure can be used [10].

The work of Gang and et al is based on the genetic method, the fitness function is modularity and distances are the chromosomes. their method does not need to specify communities number in the graph, So the number of communities as a result is obtained when the modularity measure amount is optimized [20].

Newman and et al presented a method which used a modularity function in order to quantitative evaluation of the community structure in the network. In their method, the network modularity N is defined as the fraction of all the edges in community minus the expected value of the same value in a random graph N_0, which satisfy three conditions: 1. It has the same number of vertices of N, 2. each The vertex N_0 is the same as the degree of its neighbors in N. 3. The edges are randomly positioned [7]. The modular performance of the network is defined as:

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - P_{ij}\right)\delta\left[C_i, C_j\right] \quad (1)$$

where m is the edges number in G, A_{ij} is the adjacency matrix G, the symbol i and j are the edges between nodes. P_{ij} is the expected number of edges between i and j (two considered nodes) in the model and $\delta\left[C_i, C_j\right]$ is the coronary concept. In network modularity-based strategies, the goal is to find a partition in G that has a maximum Q value which indicates the quality of the graph clustering, So the higher Q value represents better partitioning of the network. Using the modularity of network has caused the problem of finding communities to be an optimization problem which tries to find a network partition that can maximize Q value. Unfortunately, finding the maximization of Q is a NPhard problem. As a result, a meta heuristic methods is needed to find a better solution that will simultaneously guarantee logical computational costs and scalability. Existing methods based on maximizing modularity measure of a network have two major issue. The first problem is that the method actually works for small and medium sized networks in order to achieve high Q values. The second one is the limitation of clarity. Communities with a number of vertices which smaller than one threshold (which in turn depends on the number of network edges) are detected because combined optimization methods with the goal of maximizing Q value generate small groups.

Liang Bai et al presented a heuristic approach based on a based on two criteria of local importance of a node in a community and its importance concentration in all communities, They called their methods ISCD. [21].

In recent years, researchers have gradually used artificial intelligence information technology to optimize the ranking of criteria for finding the community with ideal structure. In this paper, a new method is proposed using the optimized particle swarm optimization algorithm. The proposed method uses a mutual K-nearest-neighbor graph clustering based on local mean vector as fitness function in particle swarm optimization for community.

Power System Technology
ISSN:1000-3673
Received: 06-06-2024 Revised: 15-07-2024 Accepted: 28-08-2024

2. Mutual K-nearest-neighbor clustering

Given a data set, $D = \{X_i, X_j, \dots, X_n\}$, of n points in d space, d is the dimension of each point. A mutual nearest neighbor graph, $G_{mutual} = (v, A_{mutual})$, is an undirected and weighted graph. The vertices set v contains nodes that are constructed from all the samples in D and the affinity matrix $A_{mutual} = [A_{mutual}(i,j)]_{n*n}$ is defined as

$$A_{mutual}(i,j) = \begin{cases} if & X_j \in \mathcal{N}_{p_1}(X_i) \wedge X_i \in \mathcal{N}_{p_1}(X_j) \\ 0, & otherwise \end{cases} \quad (2)$$

Where $\mathcal{N}_{p_1}(X_i)$ denotes the mutual k-nearest-neighbors of X_i, p_1 is the numbers of nearest neighbors for mutual k-nearest neighbor. When$A_{mutual}(i,j) = 1$, X_i and X_j have two connected relationship and these two points are called two connected points.

p_1 is parameter of mutual k-nearest neighbor and p_1 is used to find out the mutual p_1-nearest-neighbors of each point.

An isolated point $X_i \in D$ is defined as a point which satisfies the condition that $H(i) \leq P$, where $H(i)$ represents the number of points who have two connected relationship with X_i. P is a positive integer. The isolated set S is defined as a collection of all the isolated points.

Let $X_i \notin S$, the two point of X_i is defined as a point X_j satisfies the condition that

$$X_j \notin S \wedge A_{mutual}(i,j) = 1 \quad (3)$$

where S represents the isolated set. A dual sub-cluster is defined as a set whose element X_i and all the two points of X_i belong to it.

For clarity, the way to extract the mutual k-nearest neighbor is described below with the figure [22]:

In the Fig.1, there are points in the data samples, if we consider the parameter k to be equal to 2, then for each data sample (the red and blue square (the red square is included here as the data), select two of their closest neighborhoods and specify the direction (as in Fig.2).

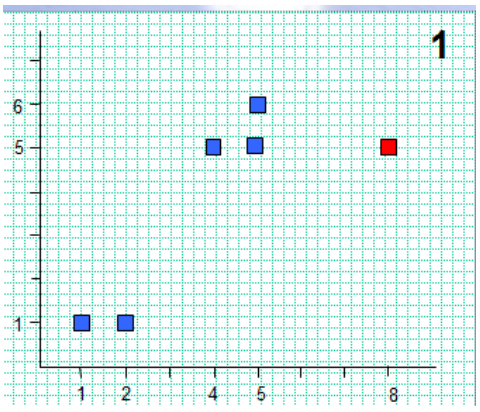

Figure 1. Data samples

Power System Technology
ISSN:1000-3673
Received: 06-06-2024 Revised: 15-07-2024 Accepted: 28-08-2024

In Figure 2, the edges obtained from k-nearest neighbor with k = 2 are shown for each data sample. Now, according to the definition of the mutual k-nearest neighbor relationship, we look at the edges that merely follow the definition of the concept of mutual k-nearest neighbor. In this figure, those data samples which an edge have been entered on it (the data sample) and an edge are out from the it (the data sample) has mutual k-nearest neighbor relationships, Indeed, an edge between two data samples has two directions indicating mutual k-nearest neighbor then for partitioning the sub-clusters, those data samples that have an edge containing two-direction relationships (i.e, edges that have two directions for entering on and outside from the data sample) as in Fig.3 is considered the sub-cluster, then after finding sub-cluster with mutual k-nearest neighbor, the edges are un-direction and the direction of the edges is eliminated after being extracted sub-clusters. The pseudo-code of the mutual k-nearest neighbor algorithm is shown in Fig.4.

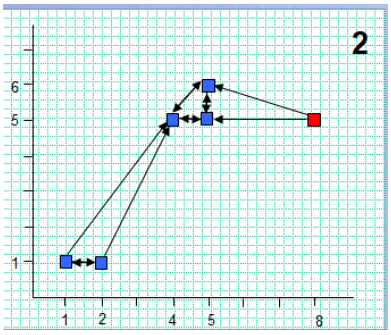

Figure 2. Finding k- nearest neighbor of each data sample (using the k = 2)

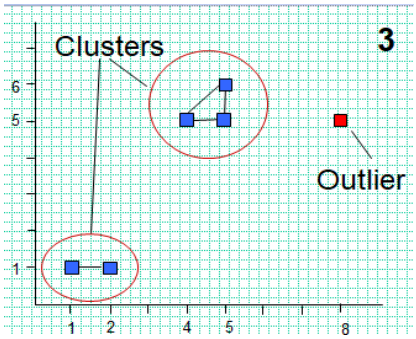

Figure 3. The clusters found using the mutual k-nearest neighbor (parameter size k =2)

Generate directed k-NN graph.
Create undirected graph as follows:
Vectors a and b are “mutual neighbors” if both links
a→ b and b→ a exist.
Change all mutual links a↔b to undirected link a—b.

Power System Technology
ISSN:1000-3673
Received: 06-06-2024 Revised: 15-07-2024 Accepted: 28-08-2024

Remove the rest.
Connected components are clusters.
Isolated vectors as outliers.

Figure 4. Pseudo-code of the mutual k-nearest neighbor algorithm

3. Proposed Method

An analysis of social networks is becoming more and more important daily. One of the most important aspects in analyzing social networks is the community detection in these networks. detecting communities in large networks has a key role in a wide range of research fields. In these networks, communities can be identified as groups of users who often interact with each other and we expect that the amount of information exchanged among members of the community is significantly higher than the amount of information exchanged between members of the community and people outside of the community. In this section, genetic optimization algorithm based on the similarity model as mutual k-nearest-neighbor clustering is used to detecting communities on social networks.

The most important problems with locally based algorithms that use evolutionary optimization (algorithms that only locally procedure to achieve solution) are not to ensure that the solution be appropriate for the problem. These problems are due to the systematic and definitive behavior of these types of algorithms. To overcome these problems, evolutionary algorithms such as genetic optimization algorithms have been introduced. The genetic optimization algorithm provides the possibility of finding the good solutions because the algorithm has population-based behavior and does not have the problem of traditional algorithms.

In this section, firstly presented a method based on local mean for improving mutual k-nearest neighbor clustering then the genetic optimization algorithm is used to optimize the improved mutual k-nearest neighbor clustering in order to detect communities in social networks.

3.1. Problem Improved mutual k-nearest neighbor clustering

Mitani et al in 2006 proposed a knn classifier based on The traditional local mean vector [24]. KNN classifier is a simple and powerful technique in pattern classification, but one major problem is that its performance severely relies on outliers. In the LMKNN method, the local mean vector is firstly computed according to all the k nearest neighbors in each class, and then the query sample is assigned to the class with the minimum euclidean distance between the query and local mean vector [24]. we improve the mutual k-nearest neighbor clustering using local mean vector which the local mean vector idea is inspired of mitani's idea. in this paper, the idea of using several k-nearest neighbor for each sample is used to find a nearest neighbor for each sample using computing local mean vector form multi k-nearest neighbor. Using local mean vector from multi k-nearest neighbor can solve problem of sensitivity to neighborhood size k and it is suitable to achieve robustness against outliers by computing local mean vector.

Power System Technology
ISSN:1000-3673
Received: 06-06-2024 Revised: 15-07-2024 Accepted: 28-08-2024

For clarify we explain detail of proposed method in follows. In order to measure the mutual relations of each samples with each other using the idea, firstly find multi k-nearest neighbor of each sample based on k parameter, Then, after extracting k-nearest neighbor sample from each sample, the mean local vector is computed, or in other words, an average of the solutions that obtained from multi k-nearest neighbor for each sample is computed. For example, first k-nearest neighbor is considered i.e, if we take k = 9, then, with respect to the value k parameter, from k = 1 to k = 9, for each k, k-nearest neighbor of each sample is measured, in the other words for parameter k with size 9 (k= 9) k-nearest neighbor, i.e. k = 1, k = 2, k = 3, k = 4, k = 5, k = 6, k = 7, k = 8, and k = 9 of each sample is computed, then for each sample, mean of all computed k-nearest neighbor (k=1,..,9). The steps and pseudocode of the proposed method is as follows. In the following code, k indicates the number of nearest neighbor (neighbor size parameter).

Select k nearest neighbor from training sample Tr_j in each sample n_j, with Euclidean distance measure, the distance measure is computed using $d(x, y_i^{NN}) = \sqrt{(x - y_i^{NN})^T (x - y_i^{NN})}$, after determining k-nearest neighbor form data sample n_j with $NN_{n_j}^k(x) = \frac{1}{k}\sum_{i=1}^{k} y_i^{NN}$, then all of resulted distances is sorted in ascending,

Compute a local mean vector, i.e. $m_{n_j}^1$ using the k-nearest neighbors training sample $NN_{n_j}^k(x)$ for each sample n_j:

$$\bar{m}_{n_j}^k = \frac{1}{k}\sum_{i=1}^{k} y_{i,j}^{NN} \qquad (4)$$

Then local mean vector with test data is shown by $\bar{m}_{n_j}^1$.

For i = 1 to k

$$d(x, y_i^{NN}) = \sqrt{(x - y_i^{NN})^T (x - y_i^{NN})}$$

END

$$\bar{m}_{n_j}^1 = \frac{1}{k}\sum_{i=1}^{k} y_i^{NN}$$

Figure 5. Pseudo-code of clustering using k-nearest neighbor local mean vector.

3.2. community detection using genetic algorithm and clustering method based on mutual k-nearest neighbor graph with local mean vector

The proposed algorithm has three phases; in the first phase, the problem is modeled and initialized, so that the social network is encoded with a string of integers as chromosome (number of chromosome is equal to vertexes number in the network which each bit in chromosome represents vertex in the network), a cluster identifier is assigned to each bit in

Power System Technology
ISSN:1000-3673
Received: 06-06-2024 Revised: 15-07-2024 Accepted: 28-08-2024

chromosome, the genetic algorithm usually uses a higher-quality population to accelerate convergence. Therefore, for each chromosome, a bit is selected randomly and assigns its cluster identifier to all its neighbors, we repeat this operation α · n times for each chromosome in the initial population where α is a parameter and α = 0.2 is used in experiments. This operation has led to local small communities, but the resulting clustering are still far from being optimal. In the second phase must determine a criterion in order to evaluate members of the population which is called fitness function. We consider accuracy of clustering method based on k-nearest neighbor graph with local mean vector as fitness function. In the third phase, we use tournament method for select the parent population in order to mating and generate offspring. In the fourth and fifth phases, crossover and mutation operators are applied. The details of the proposed genetic algorithm for community detection are described in the following,

3.2. Problem Modeling

In this section, a description of the proposed algorithm and the adopted space for network decomposition and the operators used are described. The procedure of the proposed algorithm is shown in Fig6.

Inputs:

Maximum Iteration max_G ، population size (M)، tournament size $tour$،Mating pool Sizemp، crossover propbability P_c، mutation probability Pop_m

POP← $InitializePopoluation(M)$

$POP_{parent} \leftarrow Selection(P، mp، tour)$

$POP_{child} \leftarrow Cross\&Mutation(POP_{parent}، P_c، P_m)$

$POP \leftarrow UpdataPop\ (POP، POP_{child})$

$until\ \ Stop\ Measure\ is\ met(max_G)$

$\boldsymbol{Output}: Best\ Solution$

Figure 6. The proposed method for community detection

3.2.1. Problem Representation and initialization

The G network is encoded with a string of integers:

$$x^i = \{x^1، x^2، \dots، x^n\} \quad (5)$$

Here n represents the number of vertices (nodes) in the graph network and x^i represents the cluster identifier from the set of vertices x^i, which can be any integer between 1 and n. We consider the vertices with the same cluster identifier in a community. This mode of proposed method for clustering in order to community detect does not need to specify the number of clusters in the graph because after the end of the algorithm, clusters are specified, so resulted clusters in each iteration represents communities. A graph with n vertices is partition at least n clusters, in which case each cluster contains a vertex that it is expressed with {1 2. . . n}.

Power System Technology
ISSN:1000-3673
Received: 06-06-2024 Revised: 15-07-2024 Accepted: 28-08-2024

The procedure for applying the initial population is shown in Fig7. First, for each chromosome, each bit (vertex in the network) is in a separate cluster, i.e each chromosome in the population is {1 2. . . n}. Therefore, this initial population is less compactness and each solution have a low quality. In the genetic optimization algorithm, usually a higher quality population is used to accelerate convergence. Therefore, for each chromosome, a bit is selected randomly and assigns its cluster identifier to all its neighbors, we use repeat this operation $\alpha \cdot n$ times for each chromosome in the initial population where α is a parameter and $\alpha = 0.2$ is used in experiments and n represents individual number in the chromosome. This operation has led to local small communities, but the resulting clustering are still far from being optimal.

Input: population size (M)
Generate a Population $POP\ \ x_k = \{1, 2, \dots n\}$
for each $POP\ \ x_k$ *do*
$t_{iter} \leftarrow 0$
Repeat
select a vertex v_i *randomly*
$x_k^j \leftarrow x_k^i$ *whenever* $(v_i, v_j \in E)$
$t_{iter} \leftarrow t_{iter} + 1$
until $t_{iter} = a.n$
end for
Output: *Best POP*

Figure 7. pseudo-code of initialize population

3.2.2. Objective Function

In order to evaluate the individuals within the population, a criterion must be defined for each chromosome, for this aim, we assign a value as fitness (or cost) of individual to each chromosome based on goodness measure which compute using fitness (or cost) function.
Therefore, in the objective function, the function must be maximized or minimized (fitness or cost function respectively), depending on the type of problem. The aim in this proposed method is maximizing the objective function. we need to determine fitness function. Therefore, we consider the fitness function to be the same as the accuracy rate of clustering method based on k-nearest neighbor graph with local mean vector and seeks to maximize it.

3.2.3. Selection

We use tournament selection [25] for select parent from population in order to mating and generate offspring. In tournament selection method, k individual is selected into population randomly and choose two individuals with best fitness among other individuals as parent in each iteration. The tournament selection also is suitable for situation when objective value has negative value. Fig.8 shows tournament selection procedure.

Power System Technology
ISSN:1000-3673
Received: 06-06-2024 Revised: 15-07-2024 Accepted: 28-08-2024

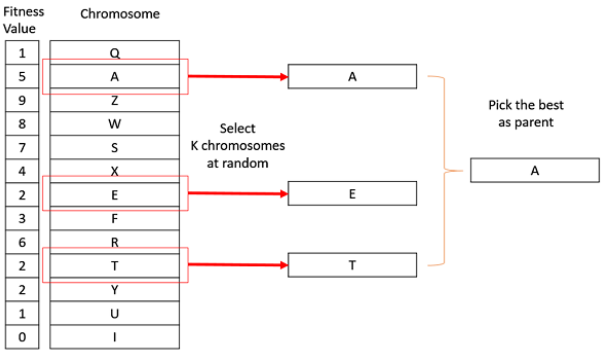

Figure 8. tournament selection

3.2.4. Crossover

We use two way crossover in [26]. Let consider the two selected chromosomes are called x_a and x_b, respectively. Firstly select a vertex v_i randomly then assign its cluster identifier (i.e., x_a^i) in the chromosome x_a and

make sure that all the vertices in this cluster of x_a are also assigned to the same cluster in the chromosome x_b (i.e. $x_b^k \leftarrow x_a^i, \forall\ k \in [k] x_a^k = x_a^i$). Simultaneously, we also assign the cluster of the vertex v_i in x_b and make sure that all the vertices in this cluster of x_b are also assigned to the same cluster in x_a (i.e. $x_a^k \leftarrow x_b^i, \forall\ k \in [k] x_b^k = x_b^i$). The procedure returns two new chromosomes x_c and x_d. The two-way crossover operation can generate descendants carrying features common to the parents, which represents the exploitative side of the crossover operator; on the other hand, the crossing over operation is exploratory, which means it can generate descendants carrying combinations of features taken from the parents.

3.2.5. Mutation

We use single- point mutation [23], in this operation, randomly a chromosome is selected for mutation, Then a single-point jump is applied to this chromosome. A vertex is randomly selected on the chromosome, then the vertex cluster randomly changes to the cluster of one of the neighbors. This action of n times is repeated on the chromosome. The mutated action that takes only the shape of the vertex neighbors can reduce search space.

4. Experiment

The proposed method is simulated using MATLAB. The proposed algorithm is studied in several steps. In the first stage, the descriptive model is calculated, then in the second stage, this model is used as the fitness function of the algorithm and then the remaining stages of the genetic optimization algorithm are performed in accordance with the proposed method in the third section.

Power System Technology
ISSN:1000-3673
Received: 06-06-2024 Revised: 15-07-2024 Accepted: 28-08-2024

We use three data are named networks which described in table1 to evaluate and validate proposed method. This data set is standard data for use in the field of community detection in the social network. The detail of databases is represented in Table.1. The Polbooks database is a network of US political books, in which the vertex (nodes) represents political books and edges represents which books are bought from Amazon. The Adjnoun database is a network that contains the proximity of names and attributes with each other, in other word, nodes represent names and attributes and edges represents the relationship between proximity between names and attributes. The polblogs database represents a network of political blogs in relation to american politics which edges in the network represents web links and communities is liberal. Table.2 contains the tuning parameters in the genetic algorithm, including population size, maximum iteration, the mutation rate and the crossover rate in the genetic algorithm. tuning parameters are determined based on trial and error experiments.

Table 1. real networks

Data	vertices	Edges
Football	115	616
Polbooks	105	441
Dolphins	62	159
Adjnoun	112	425
Polblogs	19186	16718

Table 2. tuning parameters in genetic algorithm

Parameter	Value
Pop Size	100
iter	100
Mutation rate	0.8
Crossover rate	0.1

4.1. Performance Evaluation

In this section, three benchmark networks data are used to evaluate the performance of the proposed algorithm. The network's characteristics are given in Table 1. The measure for evaluating our algorithm is detection accuracy which achieve from the normalized mutual information measure [28].

The efficiency of our algorithm has been compared with two methods, such as fast modulation maximization algorithm and mutual k-nearest neighbor clustering with local mean in section 3-1 without genetic algorithm, our algorithm and comparable algorithms are also implemented in MATLAB.

Power System Technology
ISSN:1000-3673
Received: 06-06-2024 Revised: 15-07-2024 Accepted: 28-08-2024

As mentioned above, the normalized mutual information measure is used to evaluate accuracy rate of our algorithm, so the measure is calculated from the following formula and the details of the formula are described below.

$$NMI = \frac{2\sum_i \sum_j n_{ij} \log \frac{n_{ij} n}{b_i d_j}}{-\sum_i b_i \log \frac{b_i}{n} - \sum_j d_j \log \frac{d_j}{n}} \qquad (6)$$

Where the numerator level represents mutual information and denominator level represents entropy. Table.3 represents contingency table which is used for determine mutual information in order to achieve similarity between a detection result and the true partition on each of the networks. In this case, a set of community V with n nodes and two parts $C = \{c_1, c_2, \ldots, c_s\}$ (resulted detection) and $Pi = \{pi_1, pi_2, \ldots, pi_{s'}\}$ (true partition) is considered, overlap between the two partition C and P in the possibility table (Table.3) (where in the table, n_{ij} represents common nodes number in the c_i and p_j group, or in the other words represent in form of $n_{ij} = |c_i \cap pi_j|$. Therefore, the variables n_{ij} ، b_i and d_j in the normalized mutual information formulas are the same as the values obtained from Table.3.

Table 3. Notation for the contingency table for comparing true partitions and detected partition.

$C/_{Pi}$	P_1	P_2	...	$Pi_{s'}$	result
c_1	n_{11}	n_{11}	...	$n_{1k'}$	b_1
c_2	n_{21}	n_{22}	...	$n_{2k'}$	b_2
. . .	. . .	. . .	...	. . .	. . .
c_s	n_{s1}	n_{s2}	...	$n_{ss'}$	b_s
result	d_1	d_2	...	$d_{s'}$	

If result of NMI measure on detected communities be close to true partition, the of NMI measure will increase and will be high. The proposed method in the comparison with MM in [7] and ISCD [21] yields a higher accuracy than the compared methods, but has more run time compared to the methods because the proposed method use evolutionary algorithm to calculate the optimal communities, So the run time has increased but has yielded a high degree accuracy. In high-resolution data such as the Polblogs data, which contains 19186 nodes, the high accuracy rate (72%) is reached than to compared methods, but has more runs. The reason for the high accuracy of the proposed method is use of genetic algorithm which leads optimal solution.

Power System Technology
ISSN:1000-3673
Received: 06-06-2024 Revised: 15-07-2024 Accepted: 28-08-2024

Table.4. Community detection accuracy for the proposed and compared method based on NMI measures

Data	Proposed Method	MM []	ISCD
Football	0/8519	0/3041	0/7029
Polbooks	0.7010	0.4999	0.5411
Dolphins	0/6840	0/5342	0/5184
Adjnoun	.0.6061	0.04102	0.4000
Polblogs	0/6939	0/1265	0/4402

As mentioned before, the generation's number for genetic optimization algorithm equal to 100, NMI measure values in each iteration of the genetic algorithm is shown in Fig.9 to Fig.11, as it is clear, by increasing iteration, the NMI measure has increased and this increase indicating improved accuracy with increasing iteration in our genetic algorithm. So the computed NMI measure on proposed method estimate higher value and higher difference with the comparable methods such as MM and ISCD that the resulted accuracy on proposed method is close to 100%, So we know if the NMI measure obtain higher value which close to the optimal estimation, in other word close to 100%, it represents good performance. As it is clear from Table.4, the proposed method has achieved a good performance. The difference between the proposed method and the compared methods is very considerable.

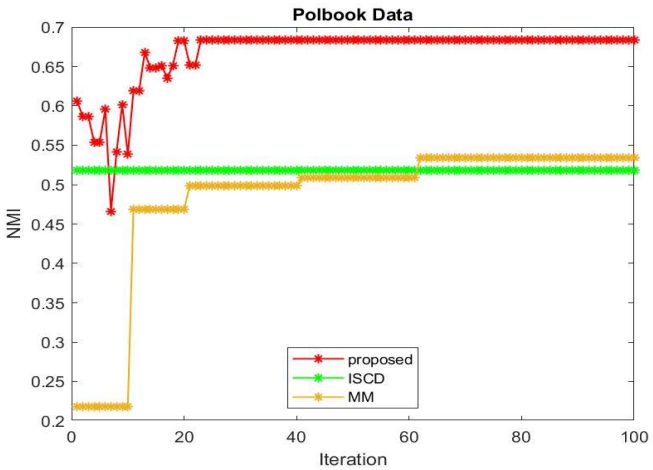

Figure 9. The resulted NMI measure on proposed method MM and ISCD on the .polbook network

Power System Technology
ISSN:1000-3673
Received: 06-06-2024 Revised: 15-07-2024 Accepted: 28-08-2024

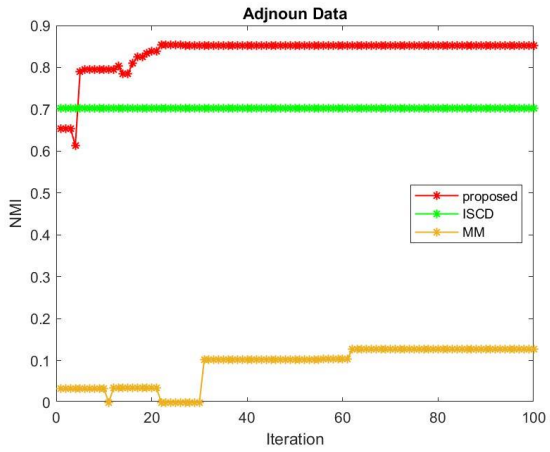

Figure 10. The resulted NMI measure on proposed method MM and ISCD on the adjnoun network.

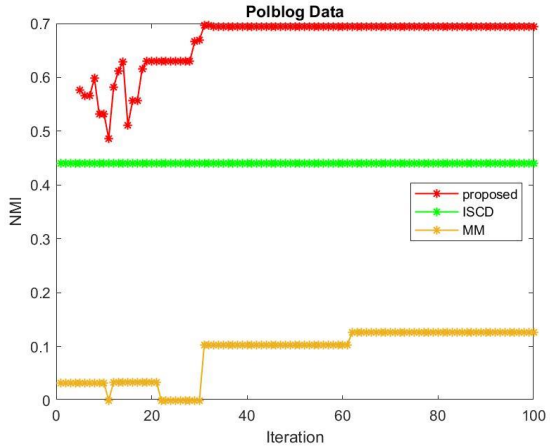

Figure 11. The resulted NMI measure on proposed method MM and ISCD on the polblogs network.

Table.5 shows the run time of the proposed method and the compared methods. The proposed method has low run time on the network with medium dimensions, but on the polblogs network, the proposed method has a high run time due to high dimensional of polblogs.

Table5. Runtime (second) on proposed and compared method

Dataset	MM [7]	ISCD [21]	Proposed Method
Polbooks	20	92	65
Adjnoun	11	78	45
Polblogs	1812	3900	2400

Power System Technology
ISSN:1000-3673
Received: 06-06-2024 Revised: 15-07-2024 Accepted: 28-08-2024

5. Conclusion

In the communication world, social networks can be considered as useful in generating and sharing beliefs and an important factor in individual and social growth. A social network is a social structure composed of various individual or organizational groups. These cross-border networks are highly has been regarded by the internet users nowadays, and every day will increase their efficiency and their adherent's interest.

The application of these spaces are mainly its potential in formation of talk and discussion place, awareness of various opinions, awareness of events and formal and informal news, participation in various political and social campaigns, improvement of friendly communication, gaining recognition of themselves and others, entertainment, and so on.

In this paper, a genetic optimization algorithm is presented to optimize the community detection model. In the proposed algorithm, a community description model was developed that considers accuracy of clustering method based on mutual k-nearest neighbor graph with local mean vector as the objective function in the genetic optimization algorithm. In the experimental analysis, the proposed algorithm is compared community detection method, i.e. MM in [7] and community detection using clustering method based on mutual k-nearest neighbor graph with local mean vector without genetic optimization. Comparative results indicate that the proposed algorithm is more effective than the other two algorithms and has better performance on large networks. In the future, we are going to improve the initialization phase of the input parameters, so that it establishes a balance between the quality of detection and the run time.

Reference

[1] M. Girvan and M. E. Newman, "Community structure in social and biological networks," Proceedings of the national academy of sciences, vol. 99, no. 12, pp. 7821-7826, 2002.
[2] G. Alexanderson, "About the cover: Euler and Königsberg's Bridges: A historical view," Bulletin of the american mathematical society, vol. 43, no. 4, pp. 567-573, 2006.
[3] S. Fortunato, "Community detection in graphs," Physics reports, vol. 486, no. 3-5, pp. 75-174, 2010.
[4] A. K. Jain and R. C. Dubes, "Algorithms for clustering data," 1988.
[5] P. Sarkar and A. W. Moore, "Dynamic social network analysis using latent space models," in Advances in Neural Information Processing Systems, 2006, pp. 1145-1152.
[6] M. E. Newman, "Modularity and community structure in networks," Proceedings of the national academy of sciences, vol. 103, no. 23, pp. 8577-8582, 2006.
[7] M. E. Newman, "Fast algorithm for detecting community structure in networks," Physical review E, vol. 69, no. 6, p. 066133, 2004.
[8] W. Ren, G. Yan, X. Liao, and L. Xiao, "Simple probabilistic algorithm for detecting community structure," Physical Review E, vol. 79, no. 3, p. 036111, 2009.

Power System Technology
ISSN:1000-3673
Received: 06-06-2024 Revised: 15-07-2024 Accepted: 28-08-2024

[9] Q. Cai, L. Ma, M. Gong, and D. Tian, "A survey on network community detection based on evolutionary computation," International Journal of Bio-Inspired Computation, vol. 8, no. 2, pp. 84-98, 2016.
[10] C. Phua, V. Lee, K. Smith, and R. Gayler, "A comprehensive survey of data mining-based fraud detection research," arXiv preprint arXiv:1009.6119, 2010.
[11] J. Huang, H. Sun, J. Han, H. Deng, Y. Sun, and Y. Liu, "Shrink: a structural clustering algorithm for detecting hierarchical communities in networks," in Proceedings of the 19th ACM international conference on Information and knowledge management, 2010, pp. 219-228: ACM.
[12] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in Advances in neural information processing systems, 2001, pp. 556-562.
[13] J. Chen and Y. Saad, "Dense subgraph extraction with application to community detection," IEEE Transactions on knowledge and data engineering, vol. 24, no. 7, pp. 1216-1230, 2012.
[14] Y. van Gennip et al., "Community detection using spectral clustering on sparse geosocial data," SIAM Journal on Applied Mathematics, vol. 73, no. 1, pp. 67-83, 2013.
[15] S. Li, H. Lou, W. Jiang, and J. Tang, "Detecting community structure via synchronous label propagation," Neurocomputing, vol. 151, pp. 1063-1075, 2015.
[16] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," Journal of statistical mechanics: theory and experiment, vol. 2008, no. 10, p. P10008, 2008.
[17] M. Rosvall and C. T. Bergstrom, "An information-theoretic framework for resolving community structure in complex networks," Proceedings of the National Academy of Sciences, vol. 104, no. 18, pp. 7327-7331, 2007.
[18] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," Proceedings of the National Academy of Sciences, vol. 105, no. 4, pp. 1118-1123, 2008.
[19] D. Zhou, E. Manavoglu, J. Li, C. L. Giles, and H. Zha, "Probabilistic models for discovering e-communities," in Proceedings of the 15th international conference on World Wide Web, 2006, pp. 173-182: ACM.
[20] Y. Li, G. Liu, and S.-y. Lao, "A genetic algorithm for community detection in complex networks," Journal of Central South University, vol. 20, no. 5, pp. 1269-1276, 2013.
[21] L. Bai, X. Cheng, J. Liang, and Y. Guo, "Fast graph clustering with a new description model for community detection," Information Sciences, vol. 388, pp. 37-47, 2017.
[22] Z. Hu and R. Bhatnagar, "Clustering algorithm based on mutual K-nearest neighbor relationships," Statistical Analysis and Data Mining: The ASA Data Science Journal, vol. 5, no. 2, pp. 100-113, 2012.

Power System Technology
ISSN:1000-3673
Received: 06-06-2024 Revised: 15-07-2024 Accepted: 28-08-2024

[23] D. E. Goldberg and J. H. Holland, "Genetic algorithms and machine learning," Machine learning, vol. 3, no. 2, pp. 95-99, 1988.
[24] Y. Mitani and Y. Hamamoto, "A local mean-based nonparametric classifier," Pattern Recognition Letters, vol. 27, no. 10, pp. 1151-1159, 2006.
[25] B. L. Miller and D. E. Goldberg, "Genetic algorithms, tournament selection, and the effects of noise," Complex systems, vol. 9, no. 3, pp. 193-212, 1995.
[26] M. Tasgin, A. Herdagdelen, and H. Bingol, "Community detection in complex networks using genetic algorithms," arXiv preprint arXiv:0711.0491, 2007.
[27] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 US election: divided they blog," in Proceedings of the 3rd international workshop on Link discovery, 2005, pp. 36-43: ACM.
[28] A. F. McDaid, D. Greene, and N. Hurley, "Normalized mutual information to evaluate overlapping community finding algorithms," arXiv preprint arXiv:1110.2515, 2011.
[29] V. Krebs, "Books about us politics," unpublished, compiled by M. Newman. Retrieved from http://www-personal. umich. edu/~ mejn/netdata, 2004.