Application of community detection based on improved genetic algorithm in air pollutant propagation network

Wenjun Liu *
School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, China
*Correspondence author e-mail: 2657932014@qq.com

Abstract. In order to analyze the propagation relationship of air pollutants in the Beijing-Tianjin-Hebei region, this paper applies the complex network to the air pollutant propagation model by analyzing the airborne pollutant propagation characteristics. A new community detection algorithm based on genetic algorithm is proposed to analyze the characteristics of air pollutant propagation network. The algorithm has combined the advantages of label propagation algorithm. The concept of community structure is added to the crossover operator, and the constrained mutation operator is used to carry out the mutation operation, so as to achieve the evolution of the population. The algorithm has found out that the air pollutant propagation network does have a community structure.

1. Introduction
With the rapid development of the economy, a large amount of energy consumption has caused local air pollution problems to become more and more serious. The perennial smog in Beijing has also attracted the attention of researchers [1]. The main purpose of this paper is to establish the air pollutant propagation network based on the relevant factors of pollutant propagation through the existing regional observation network and historical monitoring data. A large complex network is formed by regarding the air-quality stations in different regions as nodes [2]. According to the community structure of complex networks, algorithms are used to find out the community structure existing in complex networks. This process is called community detection, and such algorithms are called community detection algorithms. Mining the potential community structure in complex networks through community detection algorithms plays a crucial role in further studying the topology and hierarchy of networks [3].

Genetic algorithm (GA) is a kind of global probabilistic optimization search method by simulating biological evolution mechanism. It belongs to intelligent optimization algorithm and has excellent performance in many fields. For some classical problems like the knapsack problem, graph colouring problem, travelling salesman problem, shortest path problem and other so on, researchers can never find an absolutely correct answer to these problems, but by adding some constraints, they can find a set of solutions that reach the optimal value in some evaluation indexes [4]. In this paper, modularity $Q$ is taken as the fitness function, and the idea of label propagation algorithm (LPA) is added on the basis of genetic algorithm. The random search in the algorithm is improved to the constrained search, which improves the accuracy of the algorithm [5,6].


2. Related background

2.1. Analysis of Air Pollutant Transmission Network

Air pollutants are affected by meteorological, air pollution sources and topography, and are systematic, dynamic and complex. In order to monitor air quality, air quality monitoring stations have been set up in all parts of a country. Statistical analysis of the air quality data recorded by these stations can build an air pollutant propagation network as a whole. In this paper, each monitoring site is used as a network node. When there is a maximum transmission load between two sites, an edge is generated between the two nodes. Define a complex network of air pollutants propagation as \( G = (V, E) \), where \( V \) represents the set of all air monitoring stations, and \( E \) represents the communication relationship between two air monitoring stations. The topology of complex networks is usually represented by adjacent matrix \( A = (a_{ij})_{n \times n} \), where \( n \) is the number of nodes in the network, if two air monitoring stations satisfy the definition that there is a connection between the two, \( a_{ij} = 1 \), otherwise, \( a_{ij} = 0 \).

2.2. Build An Air Pollutant Transmission Network

After extensive research, this paper makes the following assumptions about the establishment of air pollutant propagation networks:

1. If the horizontal distance between the two monitoring stations exceeds a certain threshold, there is no edge between the two stations, and the threshold is set to 50 km [7].
2. If the vertical distance between the two monitoring sites exceeds a certain threshold, there is no edge between the two sites, and the threshold is set to 100m [8].
3. According to the numerical simulation method of atmospheric pollutants, the numerical model of the diffusion and diffusion of pollutants in the air is simulated [9]. We calculate the maximum transmission load of two monitoring stations. If the maximum transfer load is reached between the two stations, an edge is generated between the corresponding two points in the network.

2.3. Improved genetic algorithm combining with label propagation algorithm (GA-LPA)

2.3.1. Encoding and initialization. In this paper, we adopt the encoding method based on digital characters. Each gene represents the label of the corresponding node, which is the most convenient and intuitive encoding method. The coding process based on digital characters is shown in Fig. 1, which shows a community division mode of the network and the corresponding community structure diagram. The topology of the network can be seen from the network diagram in the figure. Through the network topology, we can map the network nodes and labels one by one. So we can see that node \( v_1, v_2 \), and \( v_3 \) are labeled with 1, it means they belong to the same community. And node \( v_4, v_5, v_6, v_7 \) forms a community.

In order to ensure the diversity of the population, this paper adopts the initialization method of randomly assigning labels to each node to generate a population composed of \( N \) individuals, which represents a set of community partition results.

![Figure 1. Network graph and the community structure](attachment:image.png)
2.3.2. *Multipoint crossover operator based on community structure.* Two chromosomes $p_1$ and $p_2$ are randomly selected as the crossed source and target chromosomes. Randomly select a chromosome fragment from the source chromosome $p_1$ and get the corresponding segment in $p_2$. In Fig. 2, the selected crossover fragment are between $v_2$ and $v_5$. The community structure generated in $p_2$ is used to change the structure of $p_1$, so the label of node $v_2$ is changed to 2. A new chromosome has been generated as an individual in offspring population. Exchange the source and target chromosomes to generate another new chromosome.

![Figure 2. Crossover](image)

2.3.3. *Single point mutation with constraints.* Traditional mutation operator means that the genetic value of a node is changed randomly according to a certain probability, and the rule of individuals migrating to the optimal solution in the population is random. Because the gene encoding in the community structure represents the label of the node, the traditional mutation will destroy the community structure of the original solution, resulting in invalid search in the search space. Therefore, this algorithm proposes a mutation operator with constraints. In the process of mutation is shown in Fig. 3, we randomly select a chromosome to be mutated. A random node on a chromosome is selected as the mutation node. Random selection of one of the generated community structure labels in the chromosome as the new tag of the node ensures the validity of the mutation. This can ensures the effectiveness of the mutation.

![Figure 3. Mutation](image)

2.4. *The main loop of the GA-LPA algorithm* 
Given a network $N$ and the graph $G$ modelling it, GA-LPA optimizes modularity $Q$ [10]. The algorithm starts with a population initialized at random, and then applies the modified crossover and mutation operators to produce the new population. The fitness function is used to select the offspring population.
After that, we adopt the concept of modularity, introduced by Girvan and Newman to assess the quality of a partition, to select, among the solutions found, that having the highest value of modularity.

3. Application of GA-LPA on Beijing-Tianjin-Hebei air pollutant propagation network

We take the air pollutant propagation relationship in Beijing-Tianjin-Hebei region as an example [11]. We regard each air quality monitoring station in the area as a node. When there is a maximum transmission load between the two stations, one edge is generated between the two nodes in the network.

![Figure 4. The partition of GA-LPA on Beijing-Tianjin-Hebei air pollutant propagation network](image)

As is shown in Fig. 4, the network is divided into five communities and $Q = 0.373$. The result shows that the Beijing-Tianjin-Hebei air pollutant propagation network has a certain community structure. In this figure, we can see that some nodes in different areas have been divided into the same community. It means that air pollution in a region is not only affected by emissions from local sources, but also by the spread of air pollutants outside the region.

4. Conclusion

In this paper, we propose a new community detection algorithm called GA-LPA. This algorithm is inspired by GA and LPA. We proposes a new crossover operator and mutation operator based on traditional evolutionary operators. The new crossover operator adds the concept of community to the traditional crossover operator, and it has changed the community structure in the chromosome segment when crossover, which solves the problem that the traditional crossover operator will destroy the existing community structure. The new mutation operator controls the variation in the generated community labels and it avoids invalid mutations. The algorithm's execution efficiency is improved while ensuring the accuracy of the algorithm. Finally, we use the algorithm to conduct experiments on the Beijing-Tianjin-Hebei air pollutant propagation network, so as to find the pollutant propagation relationship among Beijing-Tianjin-Hebei. And we have proved that the air pollutant propagation network has the characteristics of community structure.
References

[1] Bove M C, Brotto P, Cassola F, et al. An integrated PM 2.5 source apportionment study: Positive Matrix Factorisation vs. the chemical transport model CAMx[J]. Atmospheric Environment, 2014, 94(94):274-286.

[2] Kang X L. Research on critical path and node mining of air pollutants based on complex networks [D].

[3] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proc Natl Acad Sci U S A, 2001, 99(12):7821-7826.

[4] Osaba E, Yang X S, Diaz F, et al. An improved discrete bat algorithm for symmetric and asymmetric traveling salesman problems[J]. Engineering Applications of Artificial Intelligence, 2016, 48: 59-71.

[5] Tang C, Li Y, Duan Q, et al. Discovery of complex network communities based on adaptive evolutionary bat algorithm [J]. Journal of central south university (natural science edition), 2018, v.49;No. 281 (01) : 115-123.

[6] Raghavan U N, Albert, Réka, Kumara S. Near linear time algorithm to detect community structures in large-scale networks[J]. Physical Review E, 2007, 76(3):036106.

[7] Wang F, Chen D S, Cheng S Y, et al. Identification of regional atmospheric PM10 transport pathways using HYSPLIT, MM5-CMAQ and synoptic pressure pattern analysis[J]. Environmental Modelling & Software, 2010, 25(8):927-934.

[8] Gebali F, Yassin M F. Data mining model for atmospheric pollutants from elevated point sources[C]// Conference on Information Science. ACM, 2009.

[9] Ma F M, Gao Q X, Kang N. Numerical Simulation of an Air Pollution Process in Beijing and Surrounding Areas[J]. Meteorology and Disaster Reduction Research, 2008, 31(2).

[10] Newman M E, Girvan M. Finding and evaluating community structure in networks.[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2004, 69(2):026113.

[11] Kang X L. Research on critical path and node mining of air pollutants based on complex networks [D].