Research on Eliminating Uncertain Abnormal Large Data in Network Based on Ant Colony-Genetic Algorithms

Haiting Cui
School of Sport Communication and Information Technology, Shandong Sport University, Shandong, Jinan, 250102, China
Email: haitingsd@126.com

Abstract: In order to solve problem of high miss detection rate and false detection rate in current methods of eliminating abnormal big data in network, a method under asynchronous transmission based on ant colony-genetic algorithm is proposed. Data is validated by measuring proximity and initialization parameters of ant colony are set and real-time evolution rate of offspring is counted in which pheromones are used to train ants, then chromosomes are coded to obtain fitness function of particles. The fitness function is used to calculate individual optimal value and global optimal value of particles. Then concentration of global pheromone is used to judge whether the Ant Optimization meets end condition. If it is, large uncertain data are detected and eliminated. Experiments show that the average missed detection rate is about 2.5%, and the false detection rate is low.

1. Introduction
There are many kinds of abnormal large data under asynchronous transmission in complex network environment. Detecting and eliminating these abnormal large data is a hot issue in field of network data security. Method under asynchronous transmission based on principal component analysis (PCA) determines the dimension with high contribution rate, through which, obtains correlation between anomaly and dimension features, in order to realize adaptive correction of features. But the method has high missed detection rate and false detection rate.

Nature of network uncertain abnormal data under asynchronous transmission is complex, so it is difficult to remove it. Ant colony algorithm has characteristics of positive feedback and high parallelism, but it has the problems of slow convergence, easy stagnation and easy to fall into local optimum, combining with global-search ability of genetic algorithm, it can effectively detect abnormal large data. A method of eliminating abnormal large data in network under asynchronous transmission based on ant colony-genetic algorithm is proposed, which can effectively solve the existing problems of relevant methods.

2. Principle of Eliminating Uncertain Big Data
Big data of network connection under asynchronous transmission are discretized to calculate box dimension which is regarded as the fractal dimension of network connection and the fractal box dimension of network connection at a certain time. The decision threshold is set to make difference between fractal dimension and fractal dimension of the real-time network connection. If it exceeds the set threshold, it is determined as an exception and removed. The process is as follows:

Assuming \( I(t) \) is an asynchronous transmission network connection, to time parameter \( t \), network connection can be expressed as \( I(k) \). Assuming sampling interval is \( k \Delta t \), the \( \Omega \) box
dimension of network uncertain abnormal big data set under asynchronous transmission can be expressed as:

$$\Omega = \frac{1}{C} \sum \log \left( \frac{k_t \Delta t}{k_2 \Delta t} \right) T \log \left( \frac{k_1}{k_2} \right)$$

(1)

Fractal box dimension connected to a certain time $t_i$ is calculated. Point $\left(I_1^{(i)}, I_2^{(i)}, \ldots, I_n^{(i)}\right)$ is set during $[t_i-t_a, t_i]$ time period, we can get box dimension as $\dim(\Omega)$:

$$\dim(\Omega) = \frac{1}{C} \sum \log \left( \frac{k_t \Delta t}{\log(k_1) - \log(k_2)} \right)$$

(2)

So box dimensions has been obtained at all times. Difference between set $d_n$ and $d_c(i)$ can be defined as formula (3), fractal dimension can be recorded as $d_c(i)$ corresponded in real-time $t_i$.

$$p(i) = \left| d_c(i) - d_n \right| \dim(\Omega)$$

(3)

Set threshold value is $\vartheta$, assume that difference value made between $d_n$ and $d_c(i)$, $p(i)$ is bigger than $\vartheta$, if several consecutive times are abnormal, it is rejected.

However, in above process, the redundant data is not filtered, resulting in producing existence of duplicated abnormal data and associated normal data, which will lead to abnormal data detection and elimination methods can not completely and accurately remove abnormal big data, and the rate of missed detection and false detection will increase.

3. Ant Colony-Genetic Asynchronous Transmission Based Method for Eliminating Uncertain Abnormal Large Data in Network

3.1. Redundant Filtering of Network Uncertain Data under Asynchronous Transmission

In order to reduce accuracy and completeness of eliminating abnormal large data, redundant filtering of network uncertain data under asynchronous transmission is carried out, that consistency of data in asynchronous transmission network is verified by measuring proximity between data. The threshold value of redundant data is set and filtered and realization process is as follows:

Assume $a_i$ represents data measured by the network node $W_i$ and $a_{ij'}$ represents degree of proximity between the network nodes $W_i$ and $W_{j'}$ in asynchronous transmission, $a_{ij'}$ is as follows:

$$a_{ij'} = \begin{cases} 
1 & \left| a_i - a_{j'} \right| \leq \mu, i' \neq j' \\
0 & \left| a_i - a_{j'} \right| > \mu, i' = j' 
\end{cases}$$

(4)

Set similarity function between rows is $K_i$:

$$K_i = \frac{\sum_{j \neq i} a_{ij'}}{N-1}, i' \neq j'$$

(5)

Relevant data in network computing data can be obtained under asynchronous transmission. Assume $\nu$ respresent threshold value, Similar data $K_i > \nu$ is determined as data to be filtered and
recorded \( Q \) as a collection. To sum up, collection (6) is used to remove set \( Q \):

\[ U_{ij}^e = \frac{G(\kappa)}{Q \cdot K_e} \]  

(6)

3.2. Abnormal Data Elimination Based on Ant Colony-Genetic Algorithms

The initialization parameters of ant colony are set, and the real-time evolution rate of offspring is counted. Pheromones are used to train the ants, and then chromosomes are coded to obtain the fitness function of particles, which is used to calculate the individual optimal value and global optimal value of particles. The global pheromone concentration is used to judge whether the Ant Optimization meets the end condition. If so, the network uncertain abnormal data is detected and eliminated.

Step 1: Determine the number of ant colony particles based on uncertain data size of the network under asynchronous transmission, set is as \( \text{Chrom}^i \), initialize particle and set the particle to two-dimensional with the crossover rate is \( p_c \), Variation rate is \( p_m \), the maximum evolutionary algebra is \( \max ge \), population size is \( \text{sizegrp} \), in which of iteration times, \( gen = 0 \).

Step 2: Binary decoding of \( \text{Chrom}[i] \) to get \( x_i \).

Step 3: Fitness function of genetic pheromone can be obtained by training each particle with genetic algorithm.

\[ F^* = \frac{1}{k(x,x_e)} U_{ij}^e \]  

(7)

Step 4: Select two individuals for cross-operation. According to fitness function, the individual and global optimum values of particles can be obtained.

\[ P^*_{\text{best} x_c} = (p_x^*, p_{x_2}^*, \ldots, p_{x_r}^*) \quad \text{Chrom}_{\text{best} x_c} = (b_{g_1}, b_{g_2}, \ldots, b_{g_r}) \]  

(8)

Step 5: Data preprocessing, results of the previous step are projected into two-dimensional grid, and ants are projected into two-dimensional grid, each ant is marked as a load.

Step 6: If there is only one object in the cell, mark the ant load and add one iteration number, otherwise the ant will find the next data point and release pheromone.

Step 7: As the number of iterations increases, end condition of particle optimization is judged. When the end condition is satisfied, the best particle with highest pheromone concentration is used as the basis for optimal parameter calculation of ant colony algorithm, and then the optimal network anomaly data detection model is constructed.

\[ y(x) = \text{sign} \left( \sum_{r=1}^{N} r' \cdot \sigma' k(x,x_r) \right)_{\text{best} x_c} \]  

(9)

And, \( y(x) \) represents network anomaly data detection model, which can detect network uncertain outliers under asynchronous transmission, used to remove abnormal data:

\[ J(y) = \frac{1}{2 \cdot W^* \sigma} y(x) \]  

(10)

4. Experimental Results and Analysis

The experimental environment is: ThinkPad X280 is configured as Intel (R) i5-8250U 8G 256GSSD, Windows 10 operating system. Simulation software environment is matlab 2017. Experimental data comes from network communication data and results are as follows:
Figure 1. Missing detection rate based on data mining and information extraction and Ant Colony-Genetic Algorithms

Figure 2. Comparison of error detection rate for eliminating abnormal large data

As can be seen from Fig. 1 and Fig. 2, average missed detection rate of anomaly data detection method based on data mining and information extraction is about 13.8%, and that of network uncertain anomaly data elimination method based on ant colony-genetic algorithm under asynchronous transmission is about 2.5%. Compared with the current method, the average miss detection rate of the proposed method is lower than that of the current method.

5. Concluding
In cloud environment, correlation of large data is complex and the forms of data anomalies are various. How to mine and cluster large data effectively is a necessary means to improve efficiency. Aiming at the imperfection of current methods of detecting and filtering network anomalous data, a method of eliminating network uncertain abnormal large data under asynchronous transmission based on Ant colony-genetic algorithm is proposed to achieve cleaning of network redundant data. The ability of global optimization search of genetic algorithm makes up for deficiency of local optimization of ant colony. At last detection of pheromone concentration improves the efficiency of eliminating abnormal data. Experimental results show that the method has strong practicability and reliability.

6. Reference
[1] Wu Hao Research on optimization classification method of network anomaly data based on improved SVM [J]. Science and Technology Bulletin, 2016, 32 (2): 149-153.
[2] Mihong. Network anomaly data detection model based on intrusion feature selection [J]. Modern electronic technology, 2017, 40 (12): 69-71.
[3] Zhao Xinjie. Network anomaly detection based on migration learning and D-S theory [J]. Computer application research, 2016, 33 (4): 1137-1140.
[4] Jiang, et al. [J]. Power information network traffic anomaly detection mechanism based on large data [J]. Telecommunications Science, 2017, 33 (3): 134-141.

[5] Zhou Peng. Network state anomaly detection based on data mining [J]. Journal of Jilin University: Science Edition, 2017, 55 (5): 1269-1273.