Research Article **Open Access**

Huajie Zhang*, Sen Zhang, and Marlia Mohd Hanafiah

Localization and recognition algorithm for fuzzy anomaly data in big data networks

https://doi.org/10.1515/phys-2018-0128
Received October 4, 2018; accepted November 14, 2018

Abstract: In order to accurately detect the fuzzy anomaly data existing in big data networks, it is necessary to study the localization and recognition algorithm. The current algorithms have problems related to poor noise reduction, low recognition efficiency, high energy consumption and low accuracy. A novel localization and recognition algorithm for fuzzy anomaly data in big data networks is proposed. The *multi-wavelet denoising* method is used to remove the noise signals existing in the network. The k-means algorithm is utilized for network clustering, and the association mode between nodes and the unitary linearity regression model is adopted to eliminate spatially and temporally redundant data that exist in big data networks. The similarity anomaly detection method based on multi-feature aggregation identifies fuzzy anomaly data existing in big data networks, establishes an anomaly data localization model, and completes the localization and recognition of fuzzy anomaly data. Experimental results show that the proposed method has good noise reduction, high recognition efficiency, low energy consumption and high accuracy of localization and recognition.

Keywords: Big data networks, fuzzy anomaly data, localization and recognition, signal denoising

PACS: 07.05.Kf, 89.20.Ff, 05.40.Ca

1 Introduction

Due to the interference of deployment methods, limited energy, environmental noise, internal manufacturing defects, etc. nodes in big data networks are prone to failure or abnormal readings. This results in segmentation of the network, dynamic changes in topology and packet loss when the network is congested, especially for large-scale and high-density deployment scenarios [1]. A large number of sensor nodes may be randomly distributed in complex and harsh environments with other external interferences, such as direct electromagnetic communication interference, artificial or non-human physical and chemical factors. This can cause failure and damage to the node itself, resulting in noise, errors and defects in the collected data. Inexpensive sensor nodes have very limited computing power, memory, energy, communication bandwidth and other resources; it is inevitable that they will generate unreliable or inaccurate data [2]. When fuzzy abnormal data is present in the big data networks, it is more vulnerable to external malicious attacks, such as denial-of-service attacks, black hole attacks and eavesdropping, which pose a threat to the safe operation of big data networks. This motivates the need for locating and identifying fuzzy anomaly data. The current localization and recognition algorithm of fuzzy anomaly data have problems related to poor noise reduction, low recognition efficiency, high energy consumption, and low accuracy. Thus, localization and identification algorithms need to be further studied [3].

Liu et al. [4] proposed a localization recognition algorithm for network fuzzy anomaly data based on information gain feature selection. The algorithm normalizes the network data through the preprocessor and selects important features based on the information gain dimension reduction method. The dimension of the data set is reduced, and a random classifier is used for training and prediction to locate and identify the fuzzy anomaly data. The algorithm does not effectively remove the noise signal existing in the big data networ. Zhao et al. [5] came up with a localization and recognition algorithm for network fuzzy anomaly data based on migration learning and DS theory. The algorithm uses the migration learning method to model known network attacks and considers differences between the fuzzy abnormal data. The analysis of the unknown network behavior by the trained classifier, combined with the DS evidence theory, is used to

***Corresponding Author: Huajie Zhang:** Open Institute of Education, Zhengzhou Institute of Technology, Zhengzhou, 450044, China, E-mail: TLT333@126.com
Sen Zhang: Faculty of Science, Kunming University of Science and Technology, Kunming, 650500, China
Marlia Mohd Hanafiah: Universiti Kebangsaan Malaysia UKM, 43600 Bangi, Selangor, Malaysia

detect the fuzzy anomaly data with inconsistent distribution. The algorithm is prone to error in the training process, and the obtained localization and recognition results have low accuracy. Zhou and Xiong have proposed a localization and recognition algorithm for network fuzzy anomaly data based on data mining. The algorithm first extracts the network state signal, preprocesses the signal through a wavelet transform, and extracts the characteristics of network state anomaly detection. The fuzzy anomaly data detection is modeled with an echo state network, and the parameters of this network are optimized with a genetic algorithm, so as to achieve optimal localization and recognition of the fuzzy anomaly data. The algorithm takes a long time to identify fuzzy anomaly data, and there is a problem of low recognition efficiency [6]. Liu and Li [7] developed a localization and recognition algorithm for network anomaly data based on compressed sensing. The algorithm establishes a detection model for fuzzy anomaly data, and uses compressed sensing technology to process the temperature measurement data of the lower-level detection node collected by the upper observation node. This is combined with the sparse temperature data to construct an effective sparse matrix and measurement matrix, redefining the orthogonal transform preprocessing strategy of the measurement matrix. This makes the CS observation dictionary satisfy the constraint equidistant condition and redefines the discrete spider coding mode. The spider population is continuously co-evolved to obtain the position information of non-zero elements in the sparse results. The least squares method is utilized to obtain the amplitude information of non-zero elements and the paradigm population is iteratively evolved to obtain the parameter sequence. The localization and recognition of fuzzy anomaly data are completed by detecting the correlation threshold of the parameter sequence. When the abnormal data is located and identified, there is less residual energy in the network node, and there is a problem of high energy consumption.

In summary, a localization and recognition algorithm for fuzzy anomaly data in big data networks is proposed. The specific steps are as follows:

1) The multi-wavelet variation denoising method is used to remove noise in the network.

2) The association mode between nodes and the linear regression model are utilized to eliminate the redundant data.

3) The localization and recognition of fuzzy anomaly data are completed by the similarity anomaly detection method based on multi-feature aggregation.

Experimental results and analysis verify the overall effectiveness of the localization and recognition algorithm for fuzzy anomaly data in big data networks.

2 Methods

2.1 Multi-wavelet variation denoising

The localization and recognition algorithm for fuzzy anomaly data in the big data networks removes the noise present in the network signal through multi-wavelet variation. A new threshold function is constructed to perform secondary denoising on the processed signal [8]. The specific algorithm is as follows:

Let the discrete sample sequence of the original signal $f(x)$ in the big data network be $D = S_2^{d_0} f(n)$, $W_2^{d_j} f(n)$ be the wavelet variation value of D on each scale j, and $S_2^{d_j} f(n)$ be the approximation of D on the scale j.

The basic idea of the wavelet algorithm is to decompose the signal $S_2^{d_{j-1}} f(n)$ on each scale j into $S_2^{d_j} f(n)$ and $W_2^{d_j} f(n)$ of the next scale:

$$S_2^{d_j} f = S_2^{d_{j-1}} f \cdot h_j \tag{1}$$

$$W_2^{d_j} f = S_2^{d_{j-1}} f \cdot g_j \tag{2}$$

Where, $j = 1 \sim J$, J is the best decomposition scale, and h_j and g_j represent that inserting new filters consisting of $2^j - 1$ zeros between each adjacent two coefficients in$h$ and g, respectively.

For the wavelet coefficient $W_2^{d_1} f$ on scale 1, another three-scale wavelet transform is performed, namely:

$$S_2^{d_j}(W_2^{d_1} f) = S_2^{d_{j-1}}(W_2^{d_1} f) \cdot h_j \tag{3}$$

$$W_2^{d_1}(W_2^{d_1} f) = S_2^{d_{j-1}}(W_2^{d_1} f) \cdot g_j \tag{4}$$

In the formula, $j = 1, 2, 3$. The wavelet coefficient of the second wavelet transform on the scale 1 is set as 0, that is, $W_2^{d_1}(W_2^{d_1} f) = 0$. Then the wavelet coefficients of the remaining scales are reconstructed, and this result is used as the wavelet coefficient of the first wavelet transform on the scale 1, namely:

$$W_2^{d_{j-1}}(W_2^{d_1} f) = S_2^{d_j}(W_2^{d_1} f) \cdot h_j + W_2^{d_1}(W_2^{d_1} f) \cdot g_j \tag{5}$$

The wavelet coefficients are reconstructed with the wavelet coefficients of other scales of the first wavelet transform to obtain the denoised signal in the big data network:

$$S_2^{d_{j-1}} f = S_2^{d_j} f \cdot h_j + W_2^{d_j} f \cdot g_j \tag{6}$$

Where $j = J \sim 1$. To overcome the shortcomings of the hard threshold function and the soft threshold function, a new threshold function $\hat{w}_{j,k}$ is constructed:

$$\hat{w}_{j,k} = w_{j,k} - \lambda + \frac{2\lambda}{1+e^2} \tag{7}$$

Where $w_{j,k}$ represents the wavelet coefficient of the signal decomposition and λ is the threshold. The big data network signal is denoised twice by a new threshold function to obtain the denoised big data network signal $S_2^{'d_{j-1}} f$:

$$S_2^{'d_{j-1}} f = S_2^{d_j} f \cdot h_j + W_2^{d_j} f \cdot g_j + \hat{w}_{j,k} \tag{8}$$

2.2 Redundant data removal

The localization and recognition algorithm for fuzzy anomaly data in the big data network uses k-means algorithm to cluster the network, and mines the association mode between nodes to eliminate spatially redundant data. It establishes a linear regression model in the sensor nodes to remove temporal redundant data in the big data network, so as to improve the recognition efficiency of the algorithm.

2.2.1 Clustering

The k-means algorithm is used for clustering. The sensor data is first transmitted to the cluster head and then to the base station by the cluster head, so that a large number of nodes directly transmit the sensing data to the sink node, causing excessive energy consumption and premature death of the node [9].

The k-means algorithm is a typical distance-based clustering algorithm. Euclidean distance is used as the evaluation index of similarity measurement. That is to say, the smaller the distance between two objects, the greater the similarity. Sensing nodes in the network are densely deployed, and the spatial correlation of the sensed data from similar nodes is stronger. The specific steps of the clustering algorithm are as follows:

1) Randomly select k cluster centroid nodes in the big data network.

2) Determine the cluster to which each sensor node s_i belongs. Calculate the Euclidean distance from node s_i to each cluster centroid node u_j. The node with the smallest Euclidean distance is selected as the cluster centroid node and marked as $O[j, i] = 1$, indicating that centroid node u_j is the centroid node of node s_i.

3) Recalculate the mean of each cluster.

4) Repeat the second and third step until the cluster centroid node no longer moves.

2.2.2 Spatial correlation judgment

According to the spatial change characteristics of physical phenomena, within a certain temporal range, the perceptual data collected between adjacent sensing nodes are the same or similar, or the difference is approximately constant [10]. The localization and recognition algorithm for fuzzy anomaly data in the big data network mines the association pattern between two nodes through the historical perceptual data. If the fitting error of the historical raw data sequence of the cluster head node u_i and the intra-cluster node s_j is less than the given error threshold ε, it can be determined that the intra-cluster node s_j is spatially related to the cluster head node u_i. If the fitting error of the historical raw data sequence of the cluster head node and the intra-cluster node is greater than a given error threshold, it is spatially redundant data in the big data network, which should be eliminated.

In a certain temporal range, the latest m consecutive historical sensing data points of cluster head node u_i and intra-cluster node s_j are $U_i = \{u_i(1), u_i(2), \cdots, u_i(m)\}$ and $X_j = \{x_j(1), x_j(2), \cdots, x_j(m)\}$ respectively, then the spatial correlation between nodes u_i and s_j can be determined as follows:

First: the sequence of differences formed by nodes u_i and s_j is $\Delta X_{(i,j)}$, and the expression of $\Delta X_{(i,j)}$ is as follows:

$$\Delta X_{(i,j)} = \{\Delta x_{(i,j)}(1), \Delta x_{(i,j)}(2), \cdots, \Delta x_{(i,j)}(m)\} \tag{9}$$

Where, $\Delta x_{(i,j)}(k) = u_i(k) - x_j(k)$.

Second: calculate the mean value l of raw data sequence of nodes u_i and s_j from the difference sequence $\Delta X_{(i,j)}$:

$$l = Mean(\Delta X_{(i,j)}) = \{\Delta x_{(i,j)}(1), \Delta x_{(i,j)}(2), \cdots, \Delta x_{(i,j)}(m)\}/m \tag{10}$$

Third: calculate the fitting error $Error$ of the two sequences according to the mean l:

$$Error = \sqrt{\frac{\sum_{k=1}^{m} (\Delta x_{(i,j)}(k) - l)^2}{m}} \tag{11}$$

Fourth: if the fitting error $Error$ is smaller than the given error threshold ε, it can be assumed that the sensing data of the two nodes are related, and the correlation pattern l is stored in the correlation matrix $C[i, j]$, and vice versa, which is spatially redundant data.

Fifth: repeat First - Fourth until all the correlation patterns of the intra-cluster node and the cluster head node are determined.

When the sink node receives the sensing data of the cluster head node, the sensing data of s_j is restored by using the following formula:

$$s_j = u_i(t) - l \tag{12}$$

so that the recovered error $Error$ is less than ε.

2.2.3 Temporal correlation judgment

The nodes in the big data network periodically collect data in a high-frequency manner. For the data collected by a single node, the sampling time t can be regarded as an independent variable, and the corresponding data $x_i(t)$ is used as a piecewise linear function relationship of the dependent variable [11]. For nodes that are required to send data, the localization and recognition algorithm for fuzzy anomaly data in the big data network uses a linear regression model to eliminate temporally redundant data.

It is assumed that the linear relationship between the acquisition time t of the node s_i and the data $x_i(t)$ is the regression equation:

$$x_i(t) = \beta_0 t + \beta_1 \tag{13}$$

Knowing that the data sequence of the node s_i is $X_i = \{x_i(1), x_i(2), \cdots, x_i(m)\}$, β_0 and β_1 are parameters in the unitary linear regression model that are fitted according to the least squares method. The equations for resolving these parameters are:

$$\beta_0 = \frac{\left[\sum_{k=1}^{m} t_k x_i(k) - \frac{1}{m}(\sum_{k=1}^{m} t_k)(\sum_{k=1}^{m} x_i(k))\right]}{\left[\sum_{k=1}^{m} t_t^2 - \frac{1}{m}(\sum_{k=1}^{m} t_k)^2\right]} \tag{14}$$

$$\beta_1 = \frac{1}{m}\sum_{k=1}^{m} x_i(k) - [\frac{1}{m}(\sum_{k=1}^{m} t_k)\beta_0] \tag{15}$$

The m data points collected by the node s_i are sequentially distributed along the time axis near the fitted regression line. The constructed unitary linear regression model is shown in Figure 1.

In this diagram, δ is the absolute error of the $m + 1$-th data point and the actual value of the node s_i.

The formula for calculating δ is as follows:

$$\delta = |\acute{x_i}(m+1) - x_i(m+1)| \tag{16}$$

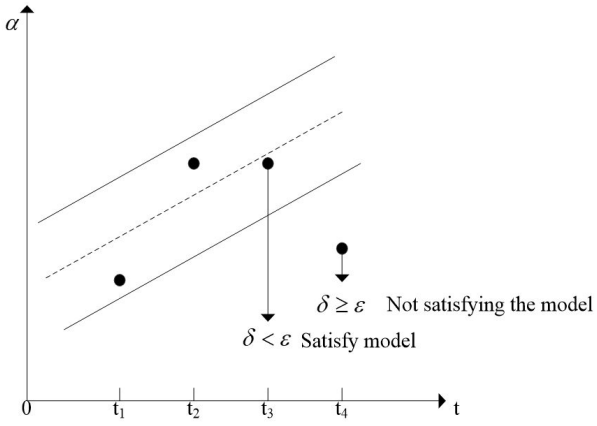


Figure 1: The unitary linear regression model

2.3 Localization and recognition of fuzzy anomaly data

2.3.1 Data similarity

Similarity is a concept in mathematics. It is used to judge the degree of difference between two data samples. The "distance" is often used to describe the degree of similarity. The larger the distance, the smaller the similarity between the two data samples [12]. A data sample can be two numbers, two sequences, or more generally, two vectors. The localization and recognition algorithm for fuzzy anomaly data in the big data network uses the Euclidean distance, as the standard to measure the similarity, and realize the localization and recognition of the fuzzy anomaly data.

Let D_e be the Euclidean distance between the two sets of samples X and Y in N-dimensional space L. The formula for calculating D_e is as follows:

$$D_e = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)} \tag{17}$$

$X = (X_1, X_2, X_3, \cdots, X_n)$, $Y = (Y_1, Y_2, Y_3, \cdots, Y_n)$.

Let D_{e1} represent the Euclidean distance between two matrix samples P and Q in $N \times M$ space S. The formula for calculating D_{e1} is as follows:

$$D_{e1} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}(P_{ij} - Q_{ij})^2} \tag{18}$$

Let $Sim(P, Q)$ represent the similarity of D_{e1}, and the formula for calculating D_{e1} is:

$$Sim(P, Q) = \frac{1}{D_{e1}} = \frac{1}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}(P_{ij} - Q_{ij})^2}} \tag{19}$$

Where P, Q are again$N \times M$-dimensional matrices, $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, m$.

Using multi-feature similarity methods to detect fuzzy anomaly data in big data networks, it is first necessary to construct a feature set of normal network states [13]. Through long-term data collection, the data is analyzed, clustered and aggregated to form a feature set per unit time and a threshold marked by time. If a uniform standard threshold is used, the periodicity of the network traffic cannot be reflected and the time stamp threshold is used, that is, the real-time traffic of a specific time period is discriminated by the threshold of a specific time, which can effectively reduce the false alarm rate [14].

The specific algorithm is as follows:

1. Initialization of feature set update times n. The initial value of n is determined by the total time T and the sampling interval time t of the training data, the expression of n is as follows:
2. $$n = \frac{T}{t} \tag{20}$$
3. The counter T is incremented by 1 each time the standard feature set needs to be updated.
4. Collect network characteristic data once every sampling time.
5. Define the $6 \times M$-dimensional real-time feature matrix I, which is used to store the feature information of the fuzzy anomaly data of the big data network.

Each row corresponds to one feature set. When there are less than m attributes in each category, it is set to 0. Each feature set is handled differently. When the source network segment exit traffic is to be stored, the cosine is required to set the column number of each network segment in the matrix [15–21].

Let the network segments A, B, and C correspond to the first, second and third column in the matrix respectively. The data of each sampling time needs to be stored according to regulations. To store the destination port traffic characteristics, it is divided into $\{(0, 100), (101, 1000), (1001, 3000), (3001, 5000), \cdots, (9000, 65535)\}$ by port number segment. Each port number segment corresponds to the columns in the matrix, in order:

$$I = \begin{bmatrix} I_{11} & I_{12} & \cdots & \cdots & \cdots & I_{1m} \\ I_{21} & I_{22} & \cdots & \cdots & \cdots & I_{2m} \\ I_{31} & I_{32} & \cdots & \cdots & \cdots & I_{3m} \\ I_{41} & I_{42} & \cdots & \cdots & \cdots & I_{4m} \\ I_{51} & I_{52} & 0 & \cdots & \cdots & 0 \\ I_{61} & I_{62} & I_{63} & 0 & \cdots & 0 \end{bmatrix} \tag{21}$$

Similarly, the $6 \times M$ standard feature set matrix S is obtained from the training data:

$$S = \begin{bmatrix} S_{11} & S_{12} & \cdots & \cdots & \cdots & S_{1m} \\ S_{21} & S_{22} & \cdots & \cdots & \cdots & S_{2m} \\ S_{31} & S_{32} & \cdots & \cdots & \cdots & S_{3m} \\ S_{41} & S_{42} & \cdots & \cdots & \cdots & S_{4m} \\ S_{51} & S_{52} & 0 & \cdots & \cdots & 0 \\ S_{61} & S_{62} & S_{63} & 0 & \cdots & 0 \end{bmatrix} \tag{22}$$

Calculate the Euclidean distance $D_e(I, S)$ of the real-time feature set matrix and the standard feature set matrix:

$$D_e(I, S) = \sqrt{\sum_{i=1}^{6} \sum_{j=1}^{m} (I_{ij} - S_{ij})^2} \tag{23}$$

Where $i = 1, 2, \cdots, 6$, $j = 1, 2, \cdots, m$. The similarity $Sim(I, S)$ of the real-time feature set matrix and the standard feature set matrix is obtained by the Euclidean distance $D_e(I, S)$:

$$Sim(I, S) = \frac{1}{D_e(I, S)} = \frac{1}{\sqrt{\sum_{i=1}^{6} \sum_{j=1}^{m} (I_{ij} - S_{ij})^2}} \tag{24}$$

If the similarity value is higher than the threshold ξ^T of this period, it is normal data, and the feature set is updated. Each attribute of the standard feature set matrix and the corresponding attribute of the real-time feature set matrix are weighted and averaged, and the updated feature set attribute is $S_{ij}(n + 1)$, and the expression is:

$$S_{ij}(n + 1) = \frac{S_{ij}(n) \cdot n + I_{ij}}{n + 1} \tag{25}$$

Let $S(n + 1)$ represent the updated feature set matrix, and the calculation formula is:

$$S(n + 1) = \frac{S(n) \cdot n + I}{n + 1} \tag{26}$$

If the similarity value is lower than the threshold ξ^T of this period, it is the fuzzy anomaly data, and the positioning model DW of the fuzzy anomaly data is constructed to complete the localization and recognition of the fuzzy anomaly data of the big data network:

$$DW = (a_1^2 + a_2^2 + \cdots + a_j^2)^r / Sim(I, S) \tag{27}$$

Where a_i represents the distance between adjacent servers in the big data network and r represents the correction factor.

3 Results

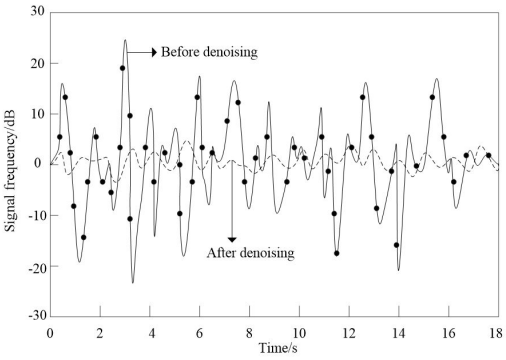
In order to verify the overall effectiveness of the localization and recognition algorithm for fuzzy anomaly data in

the big data network, it is necessary to test the localization and recognition algorithm. The operating system of this test is Windows 7.0 and the experimentation platform is Matlab. There is a large amount of noise in the big data network, which will affect the localization and recognition of fuzzy outlier data. The localization and recognition algorithm for fuzzy anomaly data in the big data network (algorithm 1), the localization and recognition algorithm for fuzzy anomaly data in the network based on information gain feature selection (algorithm 2), the localization and recognition algorithm for fuzzy anomaly data in the network based on migration learning and DS theory (algorithm 3) and the localization and recognition algorithm for fuzzy anomaly data in the network based on data mining (algorithm 4) are all tested. Four different algorithms are used to remove the noise existing in the big data network, and the denoising effects of the four different algorithms are compared. The test results are shown in Figure 2.

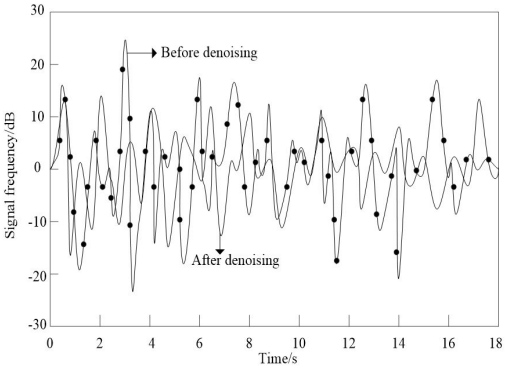
Analysis of Figure 2(a) shows that the localization and recognition algorithm for fuzzy anomaly data in the big data network is used to denoise the signal, and the signal frequency is smoother and fluctuates less frequently than the signal before the denoising. Analysis of Figure 2(b), (c) and (d) display the localization and recognition algorithm for fuzzy anomaly data in the network based on information gain feature selection, migration learning and DS theory and data mining applied to denoise the signal. The difference between the signal frequency before and after denoising is small, also the frequency after denoising fluctuates greatly. Comparing the denoising results of the four different algorithms, we can see that the localization and recognition algorithm for fuzzy anomaly data in the big data network can effectively remove the noise existing in the big data network. As the localization and recognition algorithm for fuzzy anomaly data in the big data network uses the wavelet transform method to denoise signal and construct a new threshold function to quadrate the signal, which effectively removes the noise and improves the signal-to-noise ratio.

All of the algorithms were tested, and the time used by four different algorithms to identify fuzzy anomaly data are compared, the test results are shown in Figure 3.

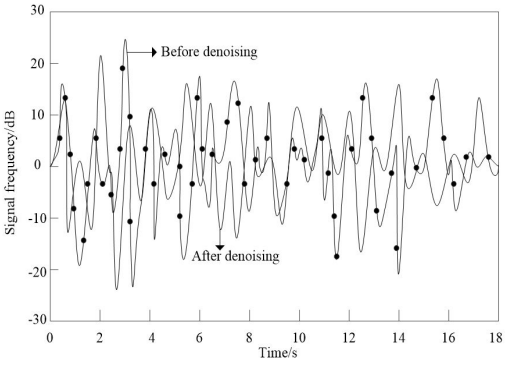
Analysis of Figure 3(a) shows that when the localization and recognition algorithms are used to identify the fuzzy anomaly data, the time used in multiple iterations is within 6 s. Figure 3(b) and (d) show that the time used in multiple iterations is as high as 8 seconds. When the network fuzzy anomaly data location recognition algorithm based on migration learning and DS theory are used to identify the fuzzy anomaly data existing in the big data network, the time used in multiple iterations is as high



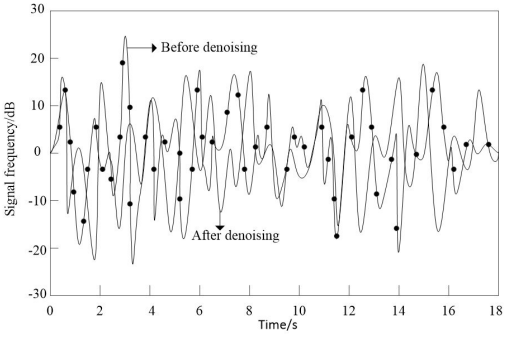
(a) Denoising effect of algorithm 1



(b) Denoising effect of algorithm 2

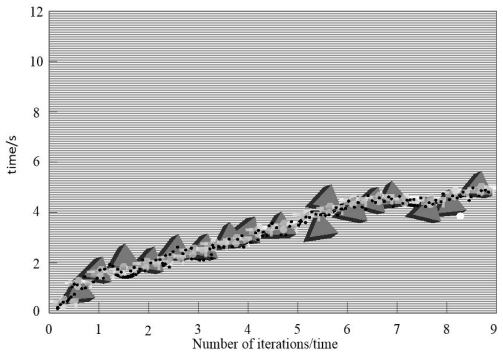


(c) Denoising effect of algorithm 3

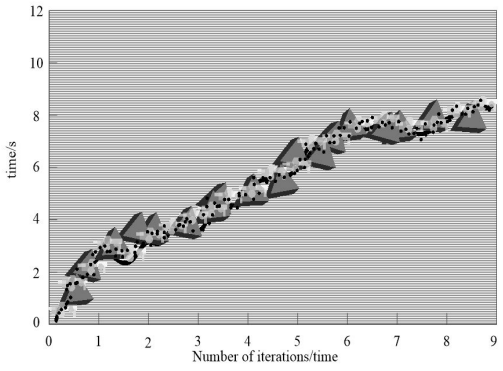


(d) Denoising effect of algorithm 4

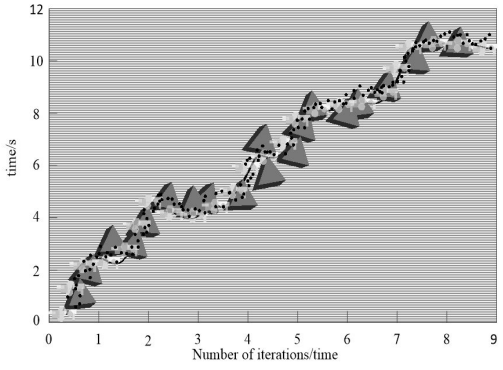
Figure 2: Denoising effects of the four different algorithms



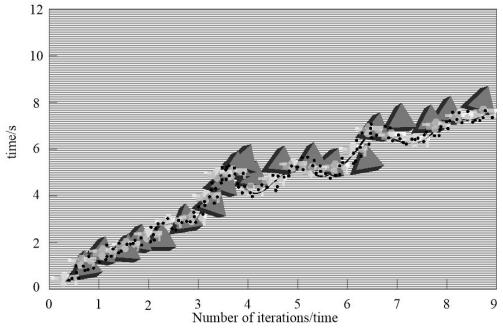
(a) Recognition time of algorithm 1



(b) Recognition time of algorithm 2



(c) Recognition time of algorithm 3



(d) Recognition time of algorithm 4

Figure 3: Recognition time of the four different algorithms

as 11 seconds. Comparing the test results of the four different algorithms, the time used by the big data network fuzzy anomaly data localization and recognition algorithm is less than the time used by the other algorithms, because it removes the spatial and the temporal redundant data existing in the big data network before localization and recognition, the time taken for processing the data is reduced and the recognition efficiency of the algorithm is improved.

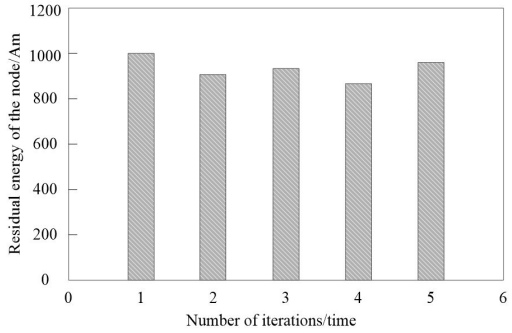
In order to further verify the overall effectiveness of the localization and recognition algorithm for fuzzy anomaly data in big data networks, the algorithms are tested in terms of energy consumption of fuzzy anomaly data localization, these results are shown in Figure 4.

Analysis of Figure 4(a) shows that after the localization and recognition algorithm is utilized for identifying and recognizing fuzzy anomaly data, the residual energy of the nodes in the network is above 800 Am. Figure 4(b) shows that when the network fuzzy anomaly data localization and recognition algorithm based on information gain feature selection is utilized to locate the fuzzy anomaly data, the residual energy of the nodes in the network is below 700 Am. Figure 4(c) shows that when the network fuzzy anomaly data localization and recognition algorithm based on migration learning and DS theory is utilized to locate the fuzzy anomaly data, the residual energy of the nodes in the network is below 500 Am. It can be seen that when the network fuzzy anomaly data localization and recognition algorithm based on data mining is utilized to locate the fuzzy anomaly data, the residual energy of the nodes in the network is below 400 Am.

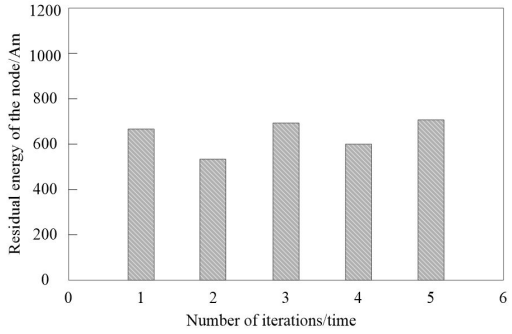
4 Discussion

τ is the correction factor in the process of fuzzy data anomaly recognition in big data network. When the correction factor τ is in the interval [2–4], the accuracy of the localization and recognition of the big data network fuzzy anomaly data is high. The result is shown in Figure 5.

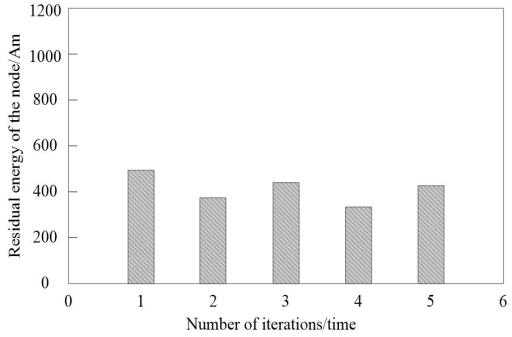
Analysis of Figure 5(a) shows that when the correction factor is used in the interval [0, 2], the accuracy of the fuzzy anomaly data localization and recognition algorithm fluctuates between 20%-50%. It can be seen from Figure 5(b) that when the correction factor is used in the interval [2, 4], the accuracy is maintained at around 80%. It can be seen from Figure 5(c) that when the correction factor is taken in the interval [4, 6], the accuracy fluctuates between 5%-35%. In summary, when the correction factor is in the interval [2, 4], the accuracy of the localization and recognition is at its highest.



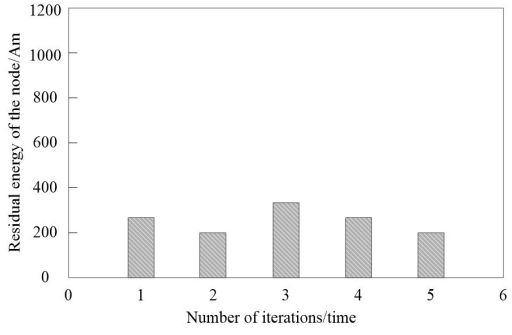
(a) Residual energy of node of algorithm 1



(b) Residual energy of node of algorithm 2

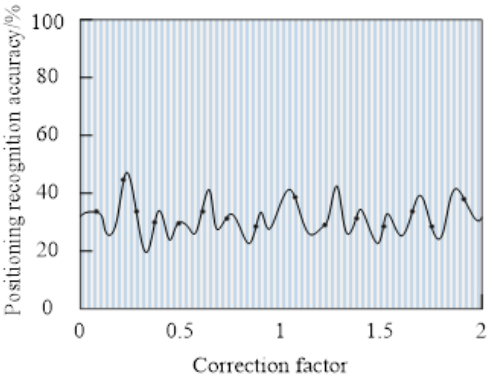


(c) Residual energy of node of algorithm 3

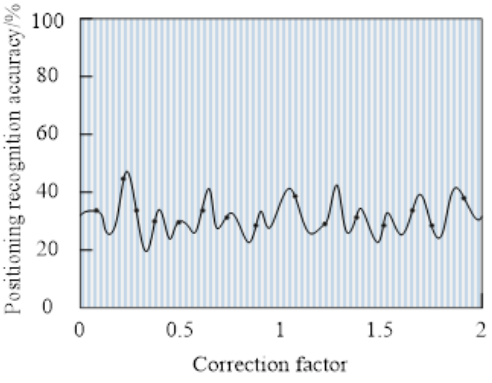


(d) Residual energy of node of algorithm 4

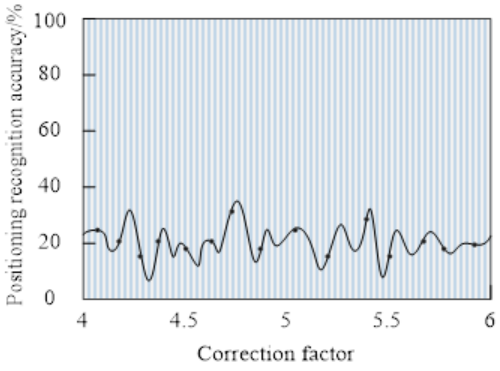
Figure 4: Residual energy of nodes of the four different algorithms



(a) Accuracy of the correction factor in the interval [0, 2]



(b) Accuracy of the correction factor in the interval [2, 4]



(c) Accuracy of the correction factor in the interval [2, 4]

Figure 5: Accuracy of correction factor in different intervals

5 Conclusion

Rapid detection of fuzzy anomaly data in big data networks, further identification of fuzzy anomaly data and reasonable response are prerequisites for ensuring the effective operation of big data networks. This is also the

frontiers of current academic and industrial research. The current localization and recognition algorithms for fuzzy anomaly data have problems of poor denoising effect, low recognition efficiency, high energy consumption and low accuracy of localization and recognition results. A new localization and recognition algorithm for fuzzy anomaly data in large data networks is proposed, which solves these problems and provides conditions for the safe operation of big data networks.

References

[1] Yan D., Dong Y., Research on Massive Network Traffic Data Analysis Based on Cloud Computing, Autom. Instrum., 2017, (9), 32-34.

[2] Liao J.J., Feng G.H. Cloud Security Vulnerability Scanning System Based on Distributed Virtual Nodes Management, J. China Acad. Electr. Inform. Technol., 2016, 11 (5), 483-489.

[3] Wang R.L., Network to Detect Abnormal Data in the Database Optimization Simulation, Comp. Simul., 2017, 34 (5), 410-413.

[4] Liu R.J., Jia B., Xin Y., Network Anomaly Detection Model Based on Information Gain Feature Selection. J. Comp. Appl., 2016, 36 (a02), 49-53.

[5] Zhao X.J., Liu Z., Sun J., New Network Anomaly Detection Using Transfer Learning and D-S Theory, Appl. Res. Comp., 2016, 33 (4), 1137-1140.

[6] Zhou P., Xiong Y.Y., Anomaly Detection of Network State Based on Data Mining, J. Jilin Univ. (Sci. Ed.), 2017, 55 (5), 1269-1273.

[7] Liu Z.Z., Li S.N., An Anomaly Detection Method for Wireless Sensor Networks Based on Compressed Sensing and GM (1, 1), J. Xi'an Jiaotong Univ., 2017, 51 (2), 40-46.

[8] Li A., An Improved Study of Abnormal Data Monitoring of Hospital Communication Networks, Electr. Design Eng., 2017, 26 (5), 165-168.

[9] Xu P., Zhang J.D., Optical Fiber Network Abnormal Data Detection Based on Improved Genetic Algorithm. Bulletin Sci. Technol., 2016, 32 (7), 163-166.

[10] Tang Y., Huang J.J., Lai M.L., Abnormal Behavior Detection Based on Integral Channel Feature Algorithm, Sci. Technol. Eng., 2016, 16 (21), 284-288.

[11] Zhang Q.C., Sun F., Wang Y.C., Under Environment of Internet Web Database Abnormal Data Detection Method Research, Comp. Meas. Contr., 2017, 25 (9), 170-173.

[12] Han S.Y., No J.G., Shin J.H., Conditional Abnormality Detection Based on Ami Data Mining, IET Gener. Trans. Distr., 2016, 10 (12), 3010-3016.

[13] Ding J., Liu Y., Zhang L., An Anomaly Detection Approach for Multiple Monitoring Data Series Based on Latent Correlation Probabilistic Model, Appl. Intel., 2016, 44 (2), 340-361.

[14] Hu X., Hu S., Huang Y., Video Anomaly Detection Using Deep Incremental Slow Feature Analysis Network, IET Comp. Vis., 2016, 10 (4), 258-265.

[15] Noble J., Adams N., Real-Time Dynamic Network Anomaly Detection. IEEE Intel. Sys., 2018, 33 (2), 5-18.

[16] Celik A., Sakin E.D., Sakin E., Seyrek A., Surface Carbon Stocks of Soil Under Pistachio Cover on Southeastern Turkey, Appl. Ecol. Envir. Res., 2017, 15(3), 747-758.

[17] Pyskunov S.O., Maksimyk Y.V., Valer V.V., Finite Element Analysis of Influence of Non-Homogenous Temperature Field On Designed Lifetime of Spatial Structural Elements Under Creep Conditions, Appl. Math. Nonlin. Sci., 2016, 1(1), 253-262.

[18] Gao W., Farahani M.R., Aslam A., Hosamani S., Distance Learning Techniques for Ontology Similarity Measuring and Ontology Mapping. Cluster Computing-the J. Net. Soft. Tools Appl., 2017, 20(2SI), 959-968.

[19] Liu Z., Peng W., Zare Y., Hui D., Rhee K.Y., Predicting the Electrical Conductivity in Polymer Carbon Nanotube Nanocomposites Based On the Volume Fractions and Resistances of the Nanoparticle, Interphase, and Tunneling Regions in Conductive Networks, Rsc Adv., 2018, 8(34), 19001-19010.

[20] López J.C.C., Quiles A.N., Bauset J.V.R., Ferragud M.D.R., Computing the Two First Probability Density Functions of the Random Cauchy-Euler Differential Equation: Study About Regular-Singular Points, Appl. Math. Nonlin. Sci., 2017, 2(1), 213-224.

[21] Wahi N., Bhatia A.K., Bhadauria S., Impact of Protozoan Vahlkampfia Sp On the Growth of Algae Chlorella Vulgaris Glamtr, J. Envir. Biol., 2018, 39(1), 109-115.