Research on Anomaly Monitoring Algorithm of Uncertain Large Data Flow Based on Artificial Intelligence

Shuang-cheng Jia(✉) and Feng-ping Yang

Alibaba Network Technology Co., Ltd., Beijing 100102, China
tomjial980@126.com

Abstract. In order to improve the monitoring ability of uncertain large data stream, an uncertain large data flow monitoring algorithm based on artificial intelligence is proposed. The collected uncertain big data flow is constructed by low dimensional feature set, and the rough set model of uncertain large data stream distribution is constructed. The fuzzy C-means clustering method is used to analyze the uncertain big data flow by fusion clustering and adaptive grid partition analysis. All the abnormal samples of large data stream are sampled and trained, and the feature quantities of association rules of uncertain large data stream are extracted. Combined with artificial intelligence method, the monitoring of uncertain large data stream is realized. The simulation results show that the method has high accuracy and good ability to resist abnormal traffic interference, and the traffic security monitoring ability of the network is improved.

Keywords: Artificial intelligence · Uncertain large data stream · Anomaly monitoring · Clustering

1 Introduction

With the development of wireless sensor network communication technology, heterogeneous directed sensor network communication technology is used to design wireless network to improve the wireless transmission ability of data. In the process of data transmission in heterogeneous directed sensor network, higher requirements are put forward for the stability of network traffic transmission [1]. The heterogeneous directed sensor network communication is affected by the interference of the transmission channel and the disturbance of inter-symbol characteristics in the data transmission, which leads to the abnormal output of the network, so it is necessary to accurately monitor the uncertain big data flow. Improve network secure transmission capacity. It has great significance to study the abnormal traffic monitoring algorithm in order to improve the security and stability of the network [2].

Combined with the category of uncertain large data stream, fuzzy clustering method is used for traffic anomaly monitoring, and the monitoring of uncertain large data stream of heterogeneous directed sensor is realized by data mining and feature

G. Gui and L. Yun (Eds.): ADHIP 2019, LNICST 302, pp. 109–117, 2019.
https://doi.org/10.1007/978-3-030-36405-2_12

extraction. The association rule feature of uncertain large data stream is extracted and filtered by anti-interference algorithm [3]. In this paper, an uncertain large data stream monitoring algorithm based on artificial intelligence is proposed. The fuzzy C-means clustering method is used to analyze the uncertain big data flow by fusion clustering and adaptive grid partition analysis. The feature extraction results of uncertain large data streams are inputted into BP neural network classifiers for data classification. Combined with big data fusion clustering method to realize uncertain big data flow monitoring, finally, the simulation experiment is carried out, which shows the superior performance of this method in improving the ability of network abnormal traffic monitoring [4].

2 Distributed Database and Feature Set Construction of Uncertain Large Data Flow

2.1 Construction of Uncertain Large Data Stream Distributed Database

To realize the monitoring of uncertain large data stream, the distributed large database model of heterogeneous directed sensor networks is constructed by using fuzzy rough clustering class method, and the nearest neighbor priority distributed information mining method is used to mine the uncertain large data stream. The adaptive association rule scheduling method is used for feature monitoring and information filtering of uncertain large data stream, and the distributed large database model of abnormal traffic in sensor network is constructed by combining correlation monitoring method [5]. The data set is vector processed, and the frequent itemsets of uncertain large data streams are calculated under the mode of uncertain frequent itemsets. The fusion analysis method of expected frequent term (EFI) and probabilistic frequent term (PFI) is adopted. The scheduling set function of uncertain large data stream is obtained as follows:

$$R_d^i(t+1) = \min\{R_s, \max\{0, R_d^i(t) + \beta(n_i - |N_i(t)|)\}\} \tag{1}$$

$$N_i(t) = \{j : ||x_j(t) - x_i(t)|| < R_d^i; l_i(t)\} \tag{2}$$

Wherein, $x_j(t)$ represents the average information entropy in the uncertain large data stream distribution data set D, describes the sample subset in the I clustering center, and $l_j(t)$ represents the sample set learned by the j generation in the process of uncertain large data flow monitoring. The output label attributes of uncertain large data streams in the i clustering center are calculated. The statistical characteristic quantity of uncertain large data stream is analyzed by split information monitoring method, and the clustering center $F(x_i, A_j(L))$, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, k$, which initializes the classification of uncertain large data stream monitoring data, is used to extract the

monitoring spectrum feature of uncertain large data stream [6]. The statistical characteristics of uncertain large data flow monitoring are obtained as follows:

$$\gamma_i = \frac{\frac{1}{w}\sum_{l=0}^{w-1}\left[x_i(k-l)-\mu_i\right]^3}{\left(\frac{1}{w}\sum_{l=0}^{w-1}\left[x_i(k-l)-\mu_i\right]^2\right)^{\frac{3}{2}}} \tag{3}$$

$$\kappa_i = \frac{\frac{1}{w}\sum_{l=0}^{w-1}\left[x_i(k-l)-\mu_i\right]^4}{\left(\frac{1}{w}\sum_{l=0}^{w-1}\left[x_i(k-l)-\mu_i\right]^2\right)^{2}} \tag{4}$$

The k uncertain large data stream monitoring impulse response function $[\delta_1, \delta_2, \cdots, \delta_N]$ becomes δ_k by extracting the association rule information of uncertain large data stream $\delta_{ik}(t)$:

$$\delta_{ik}(t) = G(V = k|U_i, \Theta(t)) \tag{5}$$

The method comprises the following steps of: carrying out fusion processing on the acquired original uncertain large data stream information, and performing beam integration processing; and the sensing information fusion model of the uncertain large data stream is expressed as follows:

$$x_m(t) = \sum_{i=1}^{I} s_i(t)e^{j\varphi_{mi}} + n_m(t),\ -p+1 \le m \le p \tag{6}$$

The collected uncertain big data flow is constructed with low dimensional feature set, and the rough set model of uncertain large data stream distribution is constructed to improve the abnormal monitoring ability of traffic flow [7].

2.2 Sensing Information Fusion for Uncertain Large Data Flow Monitoring

The adaptive regression analysis method is used to extract the statistical features of the uncertain big data flow feature set. The number of uncertain large data stream acquisition nodes is $N = n - (m-1)\tau$. The three-dimensional spectrum r_i and power spectral density k_i of uncertain large data stream monitoring are calculated. The Langevin equation is used to describe the uncertain large data flow monitoring model as follows:

$$\frac{dx}{dt} = ax - bx^2 + s(t) + \Gamma(t) \tag{7}$$

The discrete sampling and sensing information fusion tracking and recognition of the traffic sequence are carried out [8], and the abnormal statistical characteristic quantity model is obtained as follows:

$$f(x) = sgn\left\{ z\sum_{i=1}^{l_1} \alpha_i^+ y_i K(x_i, x) + \sum_{i=1}^{l_2} \alpha_i^- y_i K(x_i, x) + b \right\} \tag{8}$$

Combined with the method of scalar sequence analysis, the storage sample database model of uncertain large data stream is obtained as follows:

$$AVG_{\mathrm{X}} = \frac{1}{m \times n} \sum_{x=1}^{n} \sum_{y=1}^{m} |G_{\mathrm{X}}(x, y)| \tag{9}$$

Wherein m and n are the class number and the sampling node of the sampling sample of the uncertain large data stream, and the m, n are the classification element of the uncertain large data stream, and the frequency spectrum bandwidth of the uncertain large data stream is obtained by using the mining method of the frequent item set to obtain the frequency spectrum bandwidth of the uncertain large data stream:

$$\mathrm{sgn}(z_R^2(k) - R_{MDMMA\_R}) = \mathrm{sgn}(z_R^2(k) - \hat{e}_R^2(k)) \tag{10}$$

$$\mathrm{sgn}(z_I^2(k) - R_{MDMMA\_I}) = \mathrm{sgn}(z_I^2(k) - \hat{e}_I^2(k)) \tag{11}$$

Wherein, $\hat{e}_R^2(k)$ represents the observation sequence monitored by uncertain large data stream, $z_R^2(k)$ is the SNR of the original training set, $z_I^2(k)$ is the impulse response function of fuzzy clustering, and $\hat{e}_I^2(k)$ is the output error of the data subset. According to the above analysis, the sensing information fusion processing of network uncertainty data flow monitoring is carried out by using association rule mining method [6].

3 Optimization of Uncertain Large Data Flow Monitoring Algorithm

3.1 Feature Extraction of Uncertain Large Data Flow

Based on the rough set model of uncertain large data flow distribution, the optimal design of network uncertain large data flow monitoring is carried out. in this paper, an uncertain large data flow monitoring algorithm based on artificial intelligence is proposed. Taking a small number of sample category data as the test set, the fuzzy C-means clustering method is used to analyze the uncertain big data flow by fusion clustering and adaptive grid partition analysis [7]. In the fuzzy C-means clustering center, If the expected support degree of data element t in heterogeneous directed sensor networks is greater than the threshold θ, the attribute element monitored by

uncertain large data stream is said to be a frequent term. The classification attribute elements of all uncertain large data streams satisfy the following constraints:

$$esup^t(D) > \theta \tag{12}$$

The association characteristics of uncertain large data streams is described as: $FP\left(X_{i_j}, P_{i_j}, (sup^{k1}(D), \cdots, sup^{kf}(D)), (T_{k1}, \cdots T_{kj})\right)$, where X_{i_j} is the nth data element that appears in the first time of the uncertain large data stream arriving at the window at T_{i_j} time, and P_{i_j} is the optimal probability of output optimization training. $(sup^{k1}(D), \cdots, sup^{kf}(D))$ is a low dimensional feature set of uncertain large data streams. The correlation beamforming method is used to monitor the uncertain large data stream of the network, and the iterative formula of machine learning is obtained as follows:

$$x_O^i = x_S^i + Kd_i^{\max}(x_L^i - x_S^i) \tag{13}$$

Wherein, $K = 1/\left\|x_L^i - x_S^i\right\|$, the feature extraction results of uncertain large data streams is input into BP neural network classifiers for data classification [9], and for the calculation of $sup^t(D)$. The dynamic programming of uncertain large data flow monitoring is carried out by using big data's classified global search method [, and its calculation formula is as follows:

$$P_{i,j}^t = \begin{cases} P_{i-1,j-1}^t \times p_i + P_{i-1,j}^t \times (1 - p_i), & v_i = t \\ P_{i-1,j}^t, & v_i \neq t \end{cases} \tag{14}$$

Wherein, p_i is the probability that the distribution elements of association rules appear in the abnormal decision region i, and K is the probability of the fuzzy clustering region $P_{i,j}^t$ of the t tuples in the former S heterogeneous nodes. Sampling training of all abnormal samples of large data stream, extracting the feature quantity of association rules of uncertain large data stream, and the iterative formula of feature extraction is expressed as follows:

$$r_d^i(k+1) = \min\{r_S, \max\{0, r_d^i(k) + \beta(n_i - |N_i(k)|)\}\} \tag{15}$$

Herein, β represents the association rule feature of uncertain large data stream, and if the exception category element t satisfies the finite scheduling mode, it is called probabilistic frequent term.

$$\sum_{\omega \in PW, C^t(\omega) \geq minsup} P[\omega] > \delta \tag{16}$$

Wherein, δ is an association rule set, which realizes feature extraction of uncertain large data streams [10].

3.2 Uncertain Large Data Flow Monitoring

The decision function of uncertain large data flow monitoring is obtained by using gray scale quantitative feature analysis method:

$$U(x) = -\frac{1}{2}ax^2 + \frac{1}{4}bx^4 \tag{17}$$

In the above formula, the system parameters take the $a = 1$, $b = 1$, and the collected original heterogeneous is subjected to fusion processing to the abnormal intrusion large data information of the sensor network, and the statistical information analysis model of the uncertain large data flow monitoring is constructed, In this paper, the fuzzy cluster analysis model for uncertain large data flow monitoring is described by using the method of information fusion, and the fuzzy cluster analysis model is described as follows:

$$\omega_k = \begin{pmatrix} v_k \\ e_k \end{pmatrix} = \begin{pmatrix} x_k - f(x_{k-1}) \\ y_k - h(x_k) \end{pmatrix} \tag{18}$$

The ARMA model is used to represent the principal component characteristics of uncertain large data streams, which are described as follows:

$$x_n = a_0 + \sum_{i=1}^{M_{AR}} a_i x_{n-i} + \sum_{j=0}^{M_{MA}} b_j \eta_{n-j} \tag{19}$$

The feature quantity of association rules of uncertain large data stream is extracted, and the feature extraction results of uncertain large data stream are input into BP neural network classifier for data classification, which is analyzed comprehensively to realize the monitoring of uncertain large data stream. The monitoring steps are as follows:

Step 1. Learning coefficient of BP Neural Network for initializing uncertain large data flow monitoring $SWF = null, D = null, P_{i_j} = 0, sup^{ki}(\omega) = 0$;

Step 2. for X_{i_j}, a fuzzy clustering center point is randomly found, and the center points of all clusters monitored by uncertain large data streams are obtained.

Step 3. The statistical credential probability of abnormal network traffic monitoring is calculated according to clustering cross: P_{i_j};

Step 4. If (current window is not full), fuzzy clustering method is used to reconstruct abnormal features;

Step 5. Update the current window to reorganize the samples of uncertain large data streams, and calculate the probability distribution value of abnormal categories. $sup^{ki}(\omega)$;

Step 6. The fuzzy sample set of uncertain large data stream is calculated, and the statistical characteristic quantity is obtained by combining the cumulative probability distribution method: $Q = \sum_{i=minsup}^{num^t(D)} sup^t(D)$;

Step 7. if $Q \geq \delta$
Step 8. Adding uncertain large data flow samples of BP Learning to rough sets D;
Step 9. else
Step 10. Return frequent itemsets for uncertain large data streams D.

According to the above steps, the improved design of uncertain large data flow monitoring algorithm is realized.

4 Analysis of Simulation Experiment

In order to test the application performance of this method in the implementation of uncertain large data flow monitoring, the simulation experiment is carried out. The experiment is designed by Matlab 7 and C. The number of traffic sampling nodes in heterogeneous directed sensor networks is 200, the number of backbone nodes is 20, the number of Sink nodes is 12, and the interval between sampling points in the experiment is 5 min, uncertain large data stream injection mode is DDOS attack mode. The duration of the attack is 20 min, and the traffic anomaly features are extracted from 6 sampling points. The maximum number of iterations is 500, and the network traffic collection results are obtained from 2 to 20. The results are shown in Fig. 1.

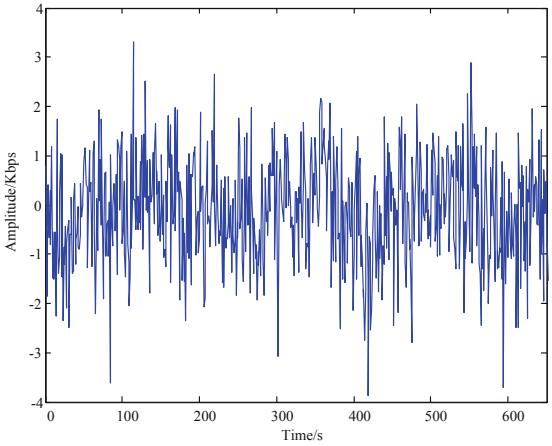

Fig. 1. Big data flow collection

The uncertainty big data flow collected in Fig. 1 is taken as the test sample to monitor the uncertain big data flow, and the monitoring output is shown in Fig. 2.

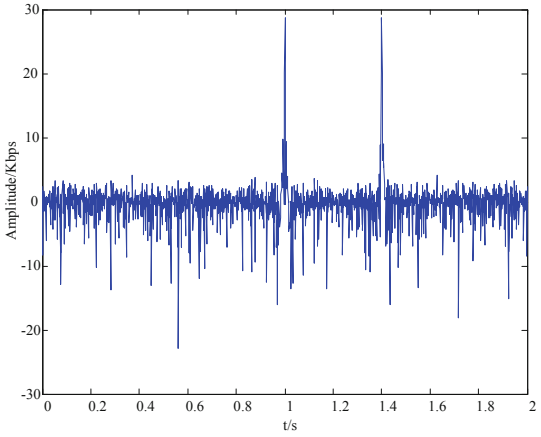

Fig. 2. Uncertain large data stream monitoring output

Figure 2 shows that the method can accurately locate the distributed frequency domain points of uncertain large data streams, and the uncertain large data streams are monitored at t = 1.04 s and t = 1.43 s, respectively. The accuracy of monitoring is good. In order to test the anti-interference and monitoring effectiveness of this method, the accuracy of monitoring uncertain large data stream is tested under different interference intensities. The results are shown in Table 1. With the increase of interference SNR, the accuracy of monitoring uncertain large data streams is increasing. When the interference intensity is 6 dB, this method can realize the integrity monitoring of uncertain large data streams. However, the accuracy of traditional methods for monitoring uncertain large data streams is lower than that of traditional methods.

Table 1. Comparison of the accuracy of different methods for monitoring large data streams (%)

Interference intensity/dB	Proposed method	Reference [4]	Reference [5]
0	95.8	89.7	78.9
4	99.9	91.2	81.1
6	1	93.4	85.6
8	1	95.6	88.7
10	1	1	89.0

5 Conclusions

In this paper, an uncertain large data flow monitoring algorithm based on artificial intelligence is proposed. The collected uncertain big data flow is constructed by low dimensional feature set, and the rough set model of uncertain large data stream distribution is constructed. The fuzzy C-means clustering method is used to analyze the uncertain big data flow by fusion clustering and adaptive grid partition analysis. All the

abnormal samples of large data stream are sampled and trained, and the feature quantities of association rules of uncertain large data stream are extracted. Combined with artificial intelligence method, the monitoring of uncertain large data stream is realized. The simulation results show that the method has high accuracy and good ability to resist abnormal traffic interference, and the traffic security monitoring ability of the network is improved. In the future, further research will be carried out on the real-time monitoring.

References

1. Wang, Z., Huang, M., et al.: Integrated algorithm based on density peaks and density-based clustering. J. Comput. Appl. **39**(2), 398–402 (2019)
2. Farnadi, G., Bach, S.H., Moens, M.F., et al.: Soft quantification in statistical relational learning. Mach. Learn. **106**(12), 1971–1991 (2017)
3. Tu, B., Chuai, R., Xu, H.: Outlier detection based on k-mean distance outlier factor for gait signal. Inf. Control **48**(1), 16–21 (2019)
4. Wei, X.S., Luo, J.H., Wu, J.: Selective convolutional descriptor aggregation for fine-grained image retrieval. IEEE Trans. Image Process. **26**(6), 2868–2881 (2017)
5. Han, D., Chen, X., Lei, Y., et al.: Real-time data analysis system based on Spark Streaming and its application. J. Comput. Appl. **37**(5), 1263–1269 (2017)
6. Hao, S.G., Zhang, L., Muhammad, G.: A union authentication protocol of cross-domain based on bilinear pairing. J. Softw. **8**(5), 1094–1100 (2013)
7. Ma, Z., Chen, W.: Friction torque calculation method of ball bearings based on rolling creepage theory. J. Mech. Eng. **53**(22), 219–224 (2017)
8. Zhou, S.B., Xu, W.X.: A novel clustering algorithm based on relative density and decision graph. Control Decis. **33**(11), 1921–1930 (2018)
9. Tu, G., Yang, X., Zhou, T.: Efficient identity-based multi-identity fully homomorphic encryption scheme. J. Comput. Appl. **39**(3), 750–755 (2019)
10. Ma, L., Zhang, T., Ma, D., Fu, Y.: Access network selection algorithm based on Markov model. Comput. Eng. **45**(5), 105–109, 115 (2019)