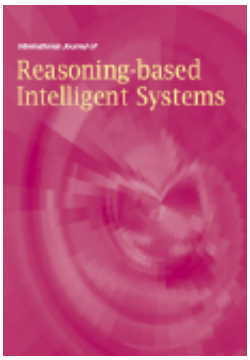


International Journal of Reasoning-based Intelligent Systems

ISSN online: 1755-0564 - ISSN print: 1755-0556
https://www.inderscience.com/ijris

Detection of redundant traffic in large-scale communication networks based on logistic regression

Xin Wen, Liyu Huang, Yin Zheng, Hailin Zhao

DOI: 10.1504/IJRIS.2024.10062136

Article History:

Received:	21 September 2022
Last revised:	07 November 2022
Accepted:	07 November 2022
Published online:	19 March 2024

Detection of redundant traffic in large-scale communication networks based on logistic regression

Xin Wen, Liyu Huang* and Yin Zheng

Guangzhou Power Supply Bureau,
Guangdong Power Grid Co., Ltd.,
Guangzhou, 510730, China
Email: underjjk@163.com
Email: 2298982155@qq.com
Email: xiaoyinyinzi@163.com
*Corresponding author

Hailin Zhao

Guangzhou Ji Neng Information Technology Co., Ltd.,
Guangzhou, 510670, China
Email: 7511029@qq.com

Abstract: In order to improve the traffic precision of network redundant traffic detection methods and reduce the time consumption of traffic classification, this paper proposes a large-scale redundant traffic detection method based on logical regression. Firstly, the logical regression architecture is analysed, and a feature extractor is constructed to extract redundant traffic features. Secondly, the weight matrix of the linear transformation between layers to be trained is obtained. Then, Gini coefficient is selected to determine the dispersion degree of redundant traffic, and redundant traffic classification function is constructed. Redundant traffic detection results are obtained through logical regression algorithm to complete network redundant traffic detection. The results show that the traffic classification time of this method is 53 ms; the precision rate is as high as 99%, which shows that the network redundant traffic detection method in this paper is effective.

Keywords: logical regression; Gini coefficient; loss function; softmax function; redundant flow detection.

Reference to this paper should be made as follows: Wen, X., Huang, L., Zheng, Y. and Zhao, H. (2024) 'Detection of redundant traffic in large-scale communication networks based on logistic regression', *Int. J. Reasoning-based Intelligent Systems*, Vol. 16, No. 1, pp.8–15.

Biographical notes: Xin Wen received his Master's in Electrical Engineering from the School of Automation of Guangdong University of Technology in 2014. He is currently an engineer in Electric Energy Measurement Center in Guangzhou Power Supply Bureau, Guang Dong Power Grid, South China Grid. His research interests include data application in power distribution network, distributed photovoltaic power generation and power demand response.

Liyu Huang received his Bachelor's in Electronics and Electrical Engineering from the School of Engineering of the University of Edinburgh and also received his PhD in Bioengineering from the School of Engineering from the University of Edinburgh in 2020. He is currently an Engineer working at the Electric Energy Measurement Center in Guangzhou Power Supply Bureau, Guang Dong Power Grid, South China Grid. His research is mainly focusing on server management and communication in the internet of things (IoT). He is also interested in the development of photovoltaic power generation in power distribution network.

Yin Zheng received her Bachelor's in Electrical Engineering and Automation in School of Electrical Engineering from the Guangdong University of Technology in 2018. She is currently an Engineer working at the Electric Energy Measurement Center in Guangzhou Power Supply Bureau, Guang Dong Power Grid, South China Grid. Her research is mainly focusing on data security in power distribution network. She is also interested in the application of electric energy data and line loss analysis.

Hailin Zhao is a Senior Engineer and also received his Master's in Computer Application Technology from the Central South University in 2005. He is currently working at the Guangzhou Ji Neng Information Technology Co., Ltd. His main research direction is electric energy measurement and electric energy data analysis and application.

1 Introduction

With the development of the internet today, network security issues are becoming more and more important, and various attacks against the network emerge in endlessly, so it is necessary to detect network abnormal traffic (Wang et al., 2021). In the network abnormal traffic detection, the traditional network abnormal traffic detection algorithm can not meet the current requirements, so machine learning has become a common method in abnormal traffic detection. In the process of using machine learning methods to detect network abnormal traffic, feature selection has become a key issue. Because of the large amount of network abnormal traffic data and various features, it may reduce the accuracy of the model and consume computing resources. Common feature selection algorithms include filtering and embedding methods. The filtering method does not consider the relationship between features and features. The disadvantage of embedding methods is that they rely too much on the final performance of the selected model (Zhang and Sun, 2022; Ding et al., 2022). In the process of redundant traffic detection, a single classifier may have some deviations. Combining multiple classifiers is a good solution. The key to traffic classification is which strategy to use to combine classifiers. In the joint strategy, Stacking has become a common joint strategy because of its good effect (Jian et al., 2020).

For this reason, relevant scholars have studied this problem. For example, Yang et al. (2020a) Applied the time-frequency domain hybrid method to the redundant traffic detection of the communication network, obtained the abnormal feature attributes of the communication network according to edge computing, classified the redundant traffic information of the communication network using discrete functions, and completed the redundant traffic detection of the communication network through the time-frequency domain hybrid feature. The classification effect of the redundant traffic information of this method is better. However, the detection accuracy of abnormal traffic in the communication network is poor. Yin et al. (2020) used the information entropy method in the communication network traffic redundancy detection, obtained the communication network traffic information through data mining method, obtained the communication network traffic redundancy characteristic attribute using data clustering method, calculated the communication network traffic redundancy data attribute weight according to the information entropy method, and judged the communication network traffic redundancy monitoring result according to the data attribute weight result, This method can improve the accuracy of network redundant traffic detection, but the detection takes a long time. Wang et al. (2020) proposed a method for detecting redundant traffic in the communication network based on the association knowledge graph. According to the high-performance processing of traffic logs by Spark computing engine, the log information of the communication network is extracted. IP, domain and files are used as the nodes of the association knowledge graph. Some node information known in the association knowledge graph is used to detect redundant traffic in the entire communication network through clustering algorithm and stain propagation algorithm, this method can effectively improve the redundant traffic detection effect, but the amount of traffic data retrieved is too large, and the traffic detection time is too long.

To solve the above problems, this paper proposes a large-scale redundant traffic detection method for communication networks based on logical regression. The specific research ideas are as follows:

First, network flow structure analysis is to analyse the network flow hierarchy of large-scale communication network, obtain the network traffic of three search levels, namely, packet level, data flow level and session level, determine the flow pcap file format, and obtain the save format of network traffic.

Secondly, feature extraction of redundant traffic in large-scale communication networks. Based on logical regression, a feature extractor for redundant traffic of large-scale communication network is constructed to extract redundant traffic features of communication network and effectively improve the detection accuracy of redundant traffic of communication network.

Then, redundant traffic detection of large-scale communication network. The Gini coefficient is selected as the impurity index to select the classification features, and the statistical characteristics of network traffic are used to determine the dispersion degree of redundant traffic, construct the classification function of network redundant traffic, use softmax to construct the loss function, obtain the detection results of network redundant traffic through the logical regression algorithm, and complete the detection of network redundant traffic.

Finally, redundant traffic precision and redundant traffic classification time are used as experimental indicators to verify the effectiveness of redundant traffic detection and draw conclusions.

2 Detection of redundant traffic in large-scale communication networks based on logical regression

2.1 Analysis of network flow structure

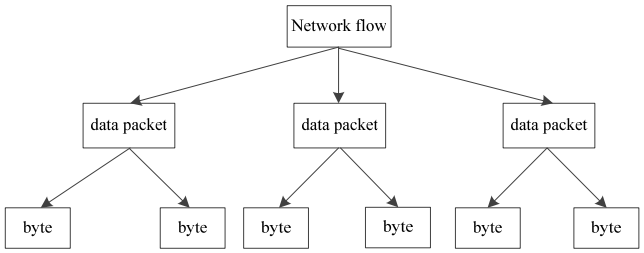
The so-called network traffic is the amount of data reflecting people's behaviour on the network. Network traffic identification refers to the application level protocol used to identify network traffic, which application the network traffic comes from, or the type of traffic carried by the network traffic (Wang et al., 2020). Due to different identification tasks or methods, the identification and search of network traffic are conducted at different levels, but in most cases it is based on the following three levels:

1 Packet level: Analyse all data packets of the entire data flow. The function analysis only considers the characteristics of the data packets, such as the size of the data packets, the distribution of the arrival time, etc.

2 Data flow level: The data flow level is mainly used to segment the captured data flow according to the source IP address, source port, target IP, target port and transport level protocol.

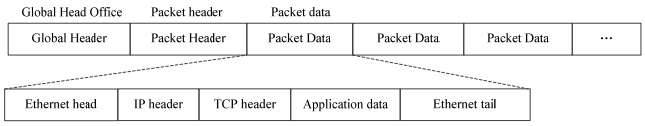
3 Session level: Divide the data packets in the data flow in the form of five tuples. As shown in Figure 1, it can be found that the corresponding objects of the source and target in this layer can be exchanged. In short, the data flow level is a collection of packets in the direction of the session flow.

Figure 1 Network flow hierarchy



As can be seen from Figure 1, the smallest unit of network flow is bytes, and multiple data packets can be combined into one network flow. In actual network traffic identification activities, network traffic data is usually saved in pcap or pcpng file format. Among them, pcap file format is the most widely used format. The dataset used in this experiment is a file in pcap format. In order to improve the accuracy of traffic detection, pcap traffic files need to be pre-processed. The file structure is shown in Figure 2. The file consists of a global header, multiple packet headers and packet data, wherein the packet header and packet data are one-to-one corresponding (Jiao et al., 2021).

Figure 2 Schematic diagram of the flow pcap file format



- Global header: The data length of this part is 24 bytes, including magic number, size and version number of the file format, accuracy of timestamp, average Greenwich Mean Time generated by the file, and the maximum byte length of the file captured.
- Packet header: The data length of this part is 16 bytes, which is mainly connected to the information overview of the received packet data in this part, including the timestamp of acquiring the details of the packet in microseconds, the length of the packet appearing on the network, and the length of the traffic data actually stored in the packet.
- Packet data: The maximum length of this part of data is 1,500 bytes, that is, the value of the maximum transmission unit, that is, the actual data transmitted at the data link layer, including user data, TCP header, IP header, Ethernet header and Ethernet tail. At this time, the format of redundant traffic in the communication network is obtained to improve the efficiency of feature extraction of redundant traffic in the communication network.

2.2 *Feature extraction of redundant traffic in large-scale communication networks*

Logic regression is a machine learning function, which is often used in classification tasks. Because of its simple model and strong interpretability, it is loved by researchers. Logical regression, also known as logistic regression analysis, is a generalised linear regression analysis model and belongs to supervised learning in machine learning. It can be considered as a linear regression function normalised by the sigmoid function (Zhou et al., 2020; Yang et al., 2020b; Zhang, 2020; Zhai et al., 2020). Logistic regression can also act as a feature extractor to find better feature representation of samples. Through the detailed analysis of the logical regression architecture, it is found that the weight matrix of the linear transformation to be trained between its layers is one of the main factors to increase the complexity and redundancy of the function. This paper assumes that the linear transformation between logical regression layers is not important, and most of the benefits of the function are due to the local smoothing of nodes. Therefore, this paper sets the output size of each layer to be consistent with the input size, and fixes the weight matrix of the linear transformation to be trained between layers as the unit matrix, but retains the nonlinear activation between layers.

At the beginning of each layer, the feature h_i of each node v_i needs to be averaged by combining the feature vectors of its neighbours in the local domain:

$$\overline{h_i^{(k)}} \leftarrow \frac{1}{d_i+1} h_i^{(k-1)} + \sum\nolimits_{j=1}^{n} \frac{a_{ij}}{\sqrt{(d_i+1)+(d_j+1)}} h_j^{(k-1)} \quad (1)$$

The update of the whole graph can be expressed as a simple matrix operation, as shown in formula (2):

$$\overline{H^{(k)}} \leftarrow SH^{(k-1)} \quad (2)$$

Intuitively, this step smoothes the hidden representation of each node locally along the edge of the graph (Yang et al., 2021). After local smoothing, in order to reduce the complexity of the model, the hidden features of redundant traffic in the large-scale communication network are linearly transformed. The feature representation output $H^{(k)}$ of the k layer of the feature extractor in this paper is:

$$H^{(k)} \leftarrow ReLu\left(\overline{H^{(k)}} I\right) \leftarrow ReLu\left(SH^{(k-1)} I\right) \quad (3)$$

wherein $ReLu()$ represents the nonlinear activation function, $H^{(0)} = X(0)$, S represents the adjacency matrix after standardisation, and I represents the feature extraction parameter. The number of layers of feature extractor K is a user-defined super parameter, so the final output of the original node features after passing through the K layer feature extractor is:

$$\hat{X} = S \cdot ReLu\left(SH^{(k-1)}I\right) I = S \cdot ReLu\left(SH^{(k-1)}\right) \quad (4)$$

In this regard, the feature extraction of redundant traffic in large-scale communication network is completed, and the obtained features are classified into redundant traffic, so as to obtain more accurate process detection results.

2.3 Detection of redundant traffic in large-scale communication networks based on logical regression

Logistic regression is an important technology in the field of artificial intelligence. Logistic regression uses high-performance computing to extract large-scale data, which solves the limitations of traditional network traffic classification in the face of user privacy and encrypted applications, and can effectively analyse the laws between data characteristics. Logic regression generally abstracts network traffic detection problem into classification problem in logic regression by using classification technology based on statistical information of packet length, duration, arrival time interval and other characteristics of network data flow during transmission.

Because the logical regression algorithm has the advantages of high efficiency, easy to explain, and less time complexity, combined with the purpose of this paper to achieve efficient encryption malicious traffic detection, we choose the logical regression as the classifier. The nodes processed by the feature extraction module will be handed over to the classification module for recognition and classification. In the given unordered training data, a classification function is created, which can identify and classify the data in a short time (Zhang, 2021). In practical use, it is found that when using the logic regression algorithm to classify binary data with small sample size (total sample size is 2,000, and 1,000 positive and negative samples are each), the Gini coefficient is selected as the impure index compared with the information entropy. Gini coefficient is an indicator to judge the fairness of income distribution according to the definition of Lorentz curve, which is a proportional value. In this paper, the Gini coefficient is introduced into the fairness of the network redundant traffic classification function to determine the impurity index and the selection information entropy. It can be predicted that if the sample size increases, the speed difference between the two will increase accordingly (Lu et al., 2021). Based on this, this paper chooses Gini coefficient as the impurity index to select the classification feature. For a dataset T, the Gini coefficient $Gini(T)$ is calculated as follows:

$$Gini(T) = \sum_{c=1}^{C} p_c\left(1-p_c\right) = 1 - \sum_{c=1}^{C} p_c^2 \quad (5)$$

where C represents the number of categories of large-scale communication network redundant traffic, and P_c represents the probability that the sample point belongs to the c category. When the probability of a large-scale communication network redundant traffic sample is P, the probability distribution of $Gini(T)$ is:

$$Gini(T) = 2P(1-p) \quad (6)$$

Intuitively, when we randomly extract two samples from dataset T, the calculated Gini coefficient represents the possibility that the two samples belong to the same category (Zhang and Wang, 2022). Therefore, the smaller $Gini(T)$ is, the greater the probability that all data in dataset T belong to the same category. By repeatedly extracting samples from the dataset to form a new dataset T to calculate Gini coefficient $Gini(T)$, we can correctly classify the samples in the dataset as much as possible.

Therefore, this paper constructs the classification function of redundant traffic in communication network through the calculated Gini coefficient above to detect redundant traffic in large-scale communication network.

Dataset $X = [X_1, X_2, \ldots, X_N]^T$ is defined as a dataset composed of network traffic samples. For each traffic sample X_i, there are n statistical characteristics, namely $X_i = (x_{i1}, x_{i2}, \ldots, x_{im})$. The category of each flow sample in the dataset is represented by vector $Y = [y_1, y_2, \ldots, y_N]^T$. When machine learning is used to detect redundant traffic in communication networks, the statistical characteristics of network traffic are used to determine the dispersion of redundant traffic. Each statistical characteristic field is defined as follows:

Let $X = [X_1, X_2, \ldots, X_N]^T$ be a set of random variables, and the maximum value $X_{\max}$ in the statistical feature satisfies that any value in the set is not greater than the maximum value, and the minimum value $X_{\min}$ satisfies that any value in the set is not less than the minimum value. The arithmetic mean represents the average value of each random variable in the set of random variables. The maximum, minimum and arithmetic mean values are shown in formulas (7), (8) and (9) below:

$$X_{\max} = \max(x) \quad (7)$$

$$X_{\min} = \min(x) \quad (8)$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \quad (9)$$

The variance in statistical characteristics is based on the arithmetic mean, reflecting the discrete degree of elements in the random variable set, and the network redundant traffic classification detection function of logical regression is obtained as shown in formula (10):

$$\delta = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{X}\right)} \quad (10)$$

The network traffic represented by statistical characteristics is used as the input of the logic regression function, and the classification of redundant traffic in large-scale communication networks based on logic regression can be realised by training the function. Next, according to the classification results of network redundant traffic, large-scale communication network redundant traffic detection is carried out.

The whole detection process is divided into three parts: the first part carries out data coding without noise and noise coding for the input data. The two coding processes are almost the same. The difference is that the noise coding adds Gaussian random noise data to each layer of the ladder network. After the input is multiplied by the weight at each layer, the batch normalisation technology is run to accelerate the training speed and improve the model accuracy, after its transformation and reconstruction, ReLu activation function is used as the input of the next layer, and iteration is performed layer by layer. After the last layer is activated with Sigmaid function, clean coding and noise coding of each layer under the original data are obtained respectively.

The second part decodes the unlabeled data using the result of noise encoding. From the last layer forward, the data of each layer is multiplied by the weight value, and then batch normalisation is performed.

The third part calculates the total loss function cost. Softmax is used to construct the supervised loss function. Its mathematical expression is shown in formulas (11)–(14):

$$\tilde{x}, \tilde{z}^{1}, \ldots, \tilde{z}^{(L)}, \tilde{y} = Encoder_{noisy}(x) \tag{11}$$

$$\tilde{x}, \tilde{z}^{(1)}, \ldots, \tilde{z}^{(L)}, \tilde{y} = Encoder_{clean}(x) \tag{12}$$

$$\hat{x}, \hat{z}^{(1)}, \ldots, \hat{z}^{(L)}, \tilde{y} = Decoder_{clean}\left(\tilde{z}^{(1)}, \ldots, \tilde{z}^{(L)}\right) \tag{13}$$

The final total loss function is:

$$\begin{aligned} Cost = &-\sum_{n=1}^{N} \log P\left(\tilde{y}(n) = y*(n) \middle| x(n)\right) \\ &+\sum_{n=N+1}^{M} \sum_{l=N+1}^{L} \lambda_l ReconstructionCost\left(z^{(l)}(n), \hat{z}^{(l)}(n)\right) \end{aligned} \tag{14}$$

Among them, x' is the data input of the sample with noise, x is the input without noise, x' is the output of decoding the encoded x, and y and y' represent the noiseless output and noisy output respectively. In the loss function training stage, the loss of supervised learning is obtained from the cross entropy loss function of y' with noise. In the loss function test stage, the sample data output y without noise is used as the evaluation of the loss function test accuracy. At this time, the total loss function output result is the redundant traffic detection result of the large-scale communication network. In this regard, the redundant traffic detection of large-scale communication network is realised by determining the storage format of traffic acquisition, extracting the characteristics of network redundant traffic, and constructing a loss function.

3 Experiment

3.1 Experimental dataset

In order to verify the effect of redundant traffic detection of large-scale communication network, MATLAB 7.0 is selected as the experimental platform, and Python 1.3.1 is used as the software framework. The selected redundant traffic detection data of communication network is shown in Table 1.

Table 1 Data distribution

Category	*.pcap number of files*	*Number of network flows*	*Number of experimental samples*
Normal	5,000	26,501	26,501
Redundancy	5,000	89,205	26,501
Total	10,000	115,706	53,002

The 53,002 processed experimental network flows came from 6,367 source IP addresses respectively (in this paper, the source and destination IP addresses of the network flow are the source and destination IP addresses of the first packet of the bidirectional network flow), and reached 1,917 destination IP addresses respectively.

In order to verify the effect of redundant traffic detection in large-scale communication networks, this paper selects the USTC-TFC2016 dataset and CIC-ISDS2017 dataset as experimental objects. The CIC-ISDS2017 dataset includes redundant circulation data such as DDos, Brute Force, XSS, SQL injection, and a total of 2,830,743 pieces of data are recorded in the pcap data table from Monday to Friday respectively; the USTC-TFC2016 dataset includes ten kinds of redundant traffic collected from the real environment, among which some files with larger size are intercepted and some smaller traffic generated by the same application is intercepted; the above datasets are detailed in Table 2.

3.2 Data pre-processing

In order to improve the detection accuracy of redundant traffic in large-scale communication network, this paper completes the data pre-processing of redundant traffic data in large-scale communication network through data cleaning and data protocol. The specific process is as follows:

First, data cleaning is completed by filling in missing values, deleting outliers, and removing redundant values. Because the dataset used in this experiment is in pcap format, all missing values in the data are filled with 0 in the experiment.

The second step is data specification. Use dimension reduction to filter the attribute subset of statistical features, remove irrelevant attributes, and construct attributes. Because in a classification task, it is often several important features in a dataset that can achieve a high accuracy rate of classification, but not all data features. The features that play a small role in classification are called redundant features. The performance of the entire classification system can be improved by deleting redundant features. Among them, the random forest feature selection algorithm can meet this requirement. The advantages of the random forest algorithm are summarised as follows:

1 It takes short time to process a large number of high-dimensional data.

2 In feature selection process, due to its randomness, it is resistant to over fitting, has high accuracy and strong robustness.

3 When constructing forests, unbiased estimates are generated.

The first 20 features with high correlation were selected from the 80 statistical features of the dataset CIC-ISD2017, and the first 20 features with high correlation were selected from the 80 statistical features of the dataset USTC-TFC2016. The specific description of the 20 features after screening is shown in Table 3.

3.3 Experimental evaluation index

The problem of redundant traffic detection in large-scale communication networks often uses accuracy ratio A_r and time consuming ratio t_m to evaluate the detection effect. Therefore, this paper also uses these two indicators in the experiments in this chapter. The calculation method of each evaluation index is as follows:

$$A_r = \frac{T_r}{T_r + T_f} \tag{15}$$

In this paper, T_r indicates that the black sample (redundant traffic) is positive, and T_f indicates that the white sample (normal traffic) is negative

$$t_m = t_2 - t_1 \pm \Delta t \tag{16}$$

In the above formula, t_1 represents the start time of redundant traffic classification, t_2 represents the end time of redundant traffic classification, and Δt represents the detection error time of redundant traffic classification.

Table 2 Introduction to datasets CIC-IDS2017 and USTC-TFC2016

Type	*USTC-TFC2016*			*CIC-IDS2017*		
	Name	*Total number of samples*	*Proportion (%)*	*Name*	*Total number of samples*	*Proportion (%)*
Normal	FTP	112,633	45	Normal	440,031	64.03
Redundancy	Geodo	51,424	15	Dos Hulk	231,073	33.63
	Miuref	14,796	6	Dos GoldEye	10,293	1.50
	Neris	49,339	5	DoS slowloris	5,796	0.84
Total		228,192	100		687,193	100

Table 3 Statistical characteristics after 20 screening

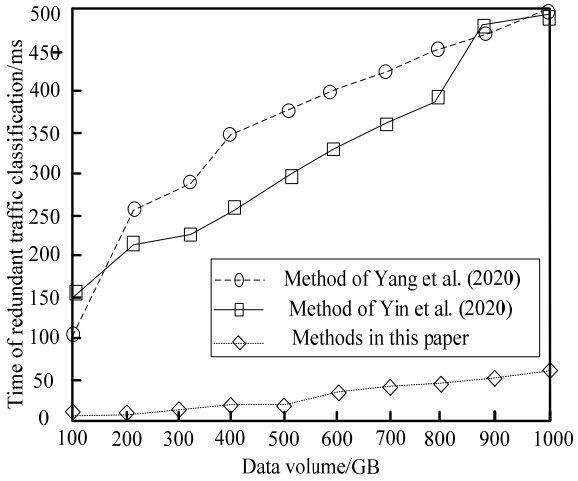
Serial no.	*Characteristic name*	*Characteristic description*
1	Packet length variance	Variance of stream length
2	Average packet size	Average packet size
3	Fwd packet length std	Standard deviation of forward packet
4	Flow bytes/s	Stream byte rate
5	Min packet length	Minimum flow length
6	Fwd header length	Length of forward head
7	PSH flag count	Number of packages with PSH
8	Init_Win_bytes_backward	Number of bytes sent by the initial window in the positive direction
9	Subflow bwd packets	Average number of packets in a subflow in the opposite direction
10	ACK flag count	Number of packages with ACKs
11	Subflow fwd bytes	Average number of bytes contained in a subflow in the positive direction
12	FIN flag count	Number of packages with FIN
13	Bwd packet length std	Standard deviation of negative packet
14	Total length of bwd packets	Total length of negative packet
15	min_seg_size_forward	Minimum slice size in the positive direction
16	URG flag count	Number of packages with URG
17	Subflow fwd packets	Average number of packets contained in a subflow in the positive direction
18	Fwd IAT std	Standard deviation between two packets sent forward
19	Fwd IAT min	Minimum time between two packets sent in the positive direction
20	Fwd IAT max	Maximum time between two packets sent in the positive direction

3.4 Result analysis

3.4.1 Time consuming for redundant traffic classification

The method in Yang et al. (2020a), Yin et al. (2020) and the method in this paper are used for time-consuming verification of redundant traffic classification. The statistical results are shown in Figure 3.

Figure 3 Time consuming of redundant traffic classification



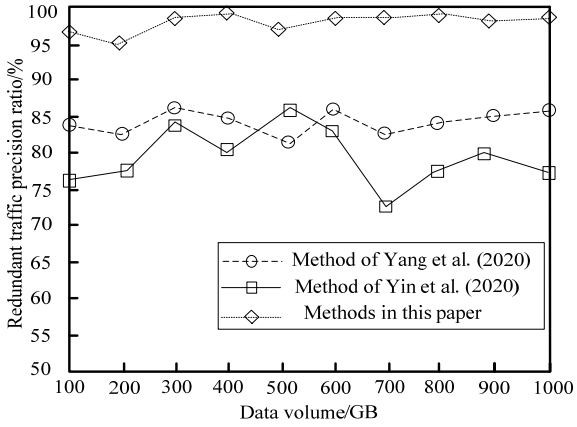
It can be seen from the analysis of Figure 3 that when the redundant traffic data volume is 100 GB, the redundant traffic classification time of Yang et al. (2020a) method is 100 ms, the redundant traffic classification time of Yin et al. (2020) method is 150 ms, and the redundant traffic classification time of this method is 2 ms; when the redundant traffic data volume is 1,000 GB, the redundant traffic classification time of Yang et al. (2020a) method is 500 ms, the redundant traffic classification time of Yin et al. (2020) method is 498 ms, and the redundant traffic classification time of this method is 53 ms; the method in this paper always has a high classification efficiency of redundant traffic, which shows that the method in this paper can classify redundant traffic in large-scale communication networks, and the classification efficiency is high. This is because the method in this paper obtains the storage format of network redundant traffic, constructs a large-scale redundant traffic feature extractor of communication network, improves the efficiency of redundant traffic feature extraction of communication network, and effectively shortens the time consumption of redundant traffic classification.

3.4.2 Redundant traffic precision

Yang et al. (2020a) method, Yin et al. (2020) method and this method are used to verify the redundancy flow precision. See Figure 4 for the statistical results.

It can be seen from the analysis of Figure 4 that when the redundant traffic data volume is 200 GB, the redundant traffic precision ratio of the Yang et al. (2020a) method is 83%, the redundant traffic precision ratio of the Yin et al. (2020) method is 72.5%, and the redundant traffic precision ratio of this method is 95%; when the redundant traffic data volume is 600 GB, the redundant traffic precision of Yang et al. (2020a) method is 87%, the redundant traffic precision of Yin et al. (2020) method is 83%, and the redundant traffic precision of this method is 99%; the method in this paper always has a high precision of redundant traffic, which indicates that the method in this paper can detect redundant traffic in large-scale communication networks with high detection accuracy. This is because the method in this paper uses the statistical characteristics of network traffic to determine the dispersion of redundant traffic, constructs the classification function of network redundant traffic, uses softmax to construct the loss function, and obtains the final network redundant traffic detection results through the logical regression algorithm, effectively improving the accuracy of network redundant traffic detection.

Figure 4 Redundant traffic accuracy



4 Conclusions

This paper proposes a method to detect redundant traffic in large-scale communication networks based on logical regression. Analyse the network flow hierarchy and determine the saving format of network traffic. Build a feature extractor for redundant traffic of large-scale communication network, complete the feature extraction of redundant traffic of communication network, build a classification detection function for redundant traffic of network, use softmax to construct a loss function, and obtain the final detection result of redundant traffic of network through logical regression algorithm. The experimental results show that:

1 This method determines the storage format of network traffic by analysing the hierarchical structure of network flow, constructs a feature extractor of network redundant traffic, improves the feature extraction effect of redundant traffic in communication network, and makes the redundant traffic classification only take 53 ms; improved classification efficiency.

2 In this method, Gini coefficient is selected as the impurity index to select the classification feature, and the statistical characteristics of network traffic are used to determine the discrete degree of redundant traffic. The classification function of network redundant traffic is constructed, and the final network redundant traffic detection result is obtained through the logical regression algorithm, making the precision of redundant traffic as high as 99%; the detection accuracy of redundant traffic in large-scale communication networks has been improved.

References

Ding, J., Liu, Y. and Liang, T. (2022) 'Network traffic anomaly detection based on feature reduction and multi-layer extreme learning machine', *Modern Electronic Technology*, Vol. 45, No. 5, pp.84–89.

Jian, S., Lu, Z., Jiang, B., Liu, Y. and Liu, B. (2022) 'Traffic anomaly detection based on hierarchical clustering method', *Information Security Research*, Vol. 6, No. 6, pp.474–481.

Jiao, L., Wang, M. and Huo, Y. (2021) 'Traffic classification and recognition method based on multi-mode deep learning', *Radio Communication Technology*, Vol. 47, No. 2, pp.215–219.

Lu, G., Tian, X. and Zhang, Y. (2021) 'Research on AMI communication intrusion detection combining KNN and optimization feature engineering', *Huadian Technology*, Vol. 43, No. 2, pp.1–8.

Wang, D., Zou, F. and Wu, Y. (2020) 'Research and implementation of network attack detection technology based on association knowledge graph', *Communication Technology*, Vol. 53, No. 12, pp.3040–3045.

Wang, J., Meng, R., Wang, Y., Zhou, L., Yuan, L. and Wang, Z. (2021) 'Power automation intelligent wireless communication network encryption traffic identification based on the oceanographic internet of things', *Microcomputer Application*, Vol. 37, No. 1, pp.140–142+151.

Yang, L., Wu, Y., Wei, D. and Pan, C. (2021) 'Satellite network traffic prediction based on space-time correlation', *Computer Engineering and Application*, Vol. 57, No. 7, pp.101–106.

Yang, T., Hou, Y., Zhao, L., Pan, H., Yuan, K. and Song, Y. (2020a) 'Detection method of abnormal flow in substation communication network based on time frequency mixed characteristics', *Power System Automation*, Vol. 44, No. 16, pp.79–86.

Yang, T., Hou, Y., Zhao, L., Pan, H., Yuan, K. and Song, Y. (2020b) 'Detection method of abnormal flow in substation communication network based on time frequency mixed characteristics', *Power System Automation*, Vol. 44, No. 16, pp.79–86.

Yin, G., Chang, C. and Li, J. (2020) 'Communication network traffic anomaly monitoring system based on information entropy', *Communication Power Technology*, Vol. 20, No. 37, pp.112–114.

Zhai, M., Zhang, X. and Zhao, B. (2020) 'Research on encrypted malicious traffic detection based on deep learning', *Journal of Network and Information Security*, Vol. 6, No. 3, pp.66–77.

Zhang, H. (2020) 'Research on cloud computing based redundant traffic suppression in communication buffer', *Information Communication*, Vol. 28, No. 7, pp.44–45.

Zhang, J. and Sun, L. (2022) 'Network anomaly detection and intelligent traffic prediction method based on deep learning', *Radio Communication Technology*, Vol. 48, No. 1, pp.81–88.

Zhang, Y. (2021) 'Research on intrusion detection method of power industrial control network based on intelligent analysis of communication flow', *Science and Technology Communication*, Vol. 13, No. 21, pp.140–142.

Zhang, Y. and Wang, Y. (2022) 'Communication information redundancy data detection method based on hierarchical aggregation', *Journal of Shanghai Institute of Electrical Engineering*, Vol. 25, No. 3, pp.182–186.

Zhou, H., Chen, C., Feng, R., Xiong, J., Pan, H. and Guo, W. (2020) 'Mobile malware traffic detection method based on value derivative GRU', *Journal of Communications*, Vol. 41, No. 1, pp.102–113.